<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.6</td>
<td>Elementary Matrices</td>
<td>92</td>
</tr>
<tr>
<td>4.7</td>
<td>Exercises</td>
<td>97</td>
</tr>
<tr>
<td>5</td>
<td>Linear Transformations</td>
<td>101</td>
</tr>
<tr>
<td>5.1</td>
<td>$\mathcal{L}(V,W)$ As A Vector Space</td>
<td>102</td>
</tr>
<tr>
<td>5.2</td>
<td>The Matrix Of A Linear Transformation</td>
<td>104</td>
</tr>
<tr>
<td>5.3</td>
<td>Rotations About A Given Vector</td>
<td>111</td>
</tr>
<tr>
<td>5.4</td>
<td>Exercises</td>
<td>112</td>
</tr>
<tr>
<td>6</td>
<td>Direct Sums And Block Diagonal Matrices</td>
<td>117</td>
</tr>
<tr>
<td>6.1</td>
<td>A Theorem Of Sylvester, Direct Sums</td>
<td>122</td>
</tr>
<tr>
<td>6.2</td>
<td>Finding The Minimum Polynomial</td>
<td>127</td>
</tr>
<tr>
<td>6.3</td>
<td>Eigenvalues And Eigenvectors Of Linear Transformations</td>
<td>129</td>
</tr>
<tr>
<td>6.4</td>
<td>A Formal Derivative, Diagonalizability</td>
<td>135</td>
</tr>
<tr>
<td>6.5</td>
<td>Exercises</td>
<td>137</td>
</tr>
<tr>
<td>7</td>
<td>Canonical Forms</td>
<td>143</td>
</tr>
<tr>
<td>7.1</td>
<td>Cyclic Sets</td>
<td>143</td>
</tr>
<tr>
<td>7.2</td>
<td>The Rational Canonical Form</td>
<td>148</td>
</tr>
<tr>
<td>7.3</td>
<td>Nilpotent Transformations</td>
<td>150</td>
</tr>
<tr>
<td>7.4</td>
<td>The Jordan Canonical Form</td>
<td>153</td>
</tr>
<tr>
<td>7.5</td>
<td>Exercises</td>
<td>156</td>
</tr>
<tr>
<td>7.6</td>
<td>Companion Matrices</td>
<td>159</td>
</tr>
<tr>
<td>7.7</td>
<td>Exercises</td>
<td>162</td>
</tr>
<tr>
<td>8</td>
<td>Determinants</td>
<td>165</td>
</tr>
<tr>
<td>8.1</td>
<td>The Function $\text{sgn}$</td>
<td>165</td>
</tr>
<tr>
<td>8.2</td>
<td>The Definition Of The Determinant</td>
<td>167</td>
</tr>
<tr>
<td>8.3</td>
<td>A Symmetric Definition</td>
<td>168</td>
</tr>
<tr>
<td>8.4</td>
<td>Basic Properties Of The Determinant</td>
<td>169</td>
</tr>
<tr>
<td>8.5</td>
<td>Expansion Using Cofactors</td>
<td>171</td>
</tr>
<tr>
<td>8.6</td>
<td>A Formula For The Inverse</td>
<td>172</td>
</tr>
<tr>
<td>8.7</td>
<td>Rank Of A Matrix</td>
<td>174</td>
</tr>
<tr>
<td>8.8</td>
<td>Summary Of Determinants</td>
<td>176</td>
</tr>
<tr>
<td>8.9</td>
<td>The Cayley Hamilton Theorem</td>
<td>177</td>
</tr>
<tr>
<td>8.10</td>
<td>Exercises</td>
<td>179</td>
</tr>
<tr>
<td>9</td>
<td>Modules And Rings</td>
<td>185</td>
</tr>
<tr>
<td>9.1</td>
<td>Integral Domains And The Ring Of Polynomials</td>
<td>185</td>
</tr>
<tr>
<td>9.2</td>
<td>Modules And Decomposition Into Cyclic Sub-Modules</td>
<td>190</td>
</tr>
<tr>
<td>9.3</td>
<td>A Direct Sum Decomposition</td>
<td>191</td>
</tr>
<tr>
<td>9.4</td>
<td>Quotients</td>
<td>194</td>
</tr>
<tr>
<td>9.5</td>
<td>Cyclic Decomposition</td>
<td>195</td>
</tr>
<tr>
<td>9.6</td>
<td>Uniqueness</td>
<td>202</td>
</tr>
<tr>
<td>9.7</td>
<td>Canonical Forms</td>
<td>204</td>
</tr>
<tr>
<td>9.8</td>
<td>Exercises</td>
<td>207</td>
</tr>
<tr>
<td>10</td>
<td>Related Topics</td>
<td>215</td>
</tr>
<tr>
<td>10.1</td>
<td>The Symmetric Polynomial Theorem</td>
<td>215</td>
</tr>
<tr>
<td>10.2</td>
<td>Transcendental Numbers</td>
<td>218</td>
</tr>
<tr>
<td>10.3</td>
<td>The Fundamental Theorem Of Algebra</td>
<td>225</td>
</tr>
<tr>
<td>10.4</td>
<td>More On Algebraic Field Extensions</td>
<td>228</td>
</tr>
<tr>
<td>CONTENTS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>14.13 Approximation In The Frobenius Norm</td>
<td>341</td>
<td></td>
</tr>
<tr>
<td>14.14 Least Squares And Singular Value Decomposition</td>
<td>343</td>
<td></td>
</tr>
<tr>
<td>14.15 The Moore Penrose Inverse</td>
<td>343</td>
<td></td>
</tr>
<tr>
<td>14.16 The Spectral Norm And The Operator Norm</td>
<td>346</td>
<td></td>
</tr>
<tr>
<td>14.17 The Positive Part Of A Hermitian Matrix</td>
<td>347</td>
<td></td>
</tr>
<tr>
<td>14.18 Exercises</td>
<td>348</td>
<td></td>
</tr>
<tr>
<td>15 Analysis Of Linear Transformations</td>
<td>353</td>
<td></td>
</tr>
<tr>
<td>15.1 The Condition Number</td>
<td>353</td>
<td></td>
</tr>
<tr>
<td>15.2 The Spectral Radius</td>
<td>354</td>
<td></td>
</tr>
<tr>
<td>15.3 Series And Sequences Of Linear Operators</td>
<td>357</td>
<td></td>
</tr>
<tr>
<td>15.4 Iterative Methods For Linear Systems</td>
<td>361</td>
<td></td>
</tr>
<tr>
<td>15.5 Theory Of Convergence</td>
<td>366</td>
<td></td>
</tr>
<tr>
<td>15.6 Exercises</td>
<td>368</td>
<td></td>
</tr>
<tr>
<td>16 Numerical Methods, Eigenvalues</td>
<td>375</td>
<td></td>
</tr>
<tr>
<td>16.1 The Power Method For Eigenvalues</td>
<td>375</td>
<td></td>
</tr>
<tr>
<td>16.1.1 The Shifted Inverse Power Method</td>
<td>378</td>
<td></td>
</tr>
<tr>
<td>16.1.2 The Explicit Description Of The Method</td>
<td>379</td>
<td></td>
</tr>
<tr>
<td>16.2 Automation With Matlab</td>
<td>379</td>
<td></td>
</tr>
<tr>
<td>16.2.1 Complex Eigenvalues</td>
<td>382</td>
<td></td>
</tr>
<tr>
<td>16.2.2 Rayleigh Quotients And Estimates for Eigenvalues</td>
<td>383</td>
<td></td>
</tr>
<tr>
<td>16.3 The QR Algorithm</td>
<td>386</td>
<td></td>
</tr>
<tr>
<td>16.3.1 Basic Properties And Definition</td>
<td>386</td>
<td></td>
</tr>
<tr>
<td>16.3.2 The Case Of Real Eigenvalues</td>
<td>389</td>
<td></td>
</tr>
<tr>
<td>16.3.3 The QR Algorithm In The General Case</td>
<td>393</td>
<td></td>
</tr>
<tr>
<td>16.3.4 Upper Hessenberg Matrices</td>
<td>397</td>
<td></td>
</tr>
<tr>
<td>16.4 Exercises</td>
<td>400</td>
<td></td>
</tr>
<tr>
<td>III Analysis Which Involves Linear Algebra</td>
<td>403</td>
<td></td>
</tr>
<tr>
<td>17 The Derivative, A Linear Transformation</td>
<td>405</td>
<td></td>
</tr>
<tr>
<td>17.1 Basic Definitions</td>
<td>405</td>
<td></td>
</tr>
<tr>
<td>17.2 The Chain Rule</td>
<td>407</td>
<td></td>
</tr>
<tr>
<td>17.3 The Matrix Of The Derivative</td>
<td>407</td>
<td></td>
</tr>
<tr>
<td>17.4 A Mean Value Inequality</td>
<td>409</td>
<td></td>
</tr>
<tr>
<td>17.5 Existence Of The Derivative, $C^1$ Functions</td>
<td>410</td>
<td></td>
</tr>
<tr>
<td>17.6 Higher Order Derivatives</td>
<td>413</td>
<td></td>
</tr>
<tr>
<td>17.7 Some Standard Notation</td>
<td>414</td>
<td></td>
</tr>
<tr>
<td>17.8 The Derivative And The Cartesian Product</td>
<td>415</td>
<td></td>
</tr>
<tr>
<td>17.9 Mixed Partial Derivatives</td>
<td>418</td>
<td></td>
</tr>
<tr>
<td>17.10 Newton's Method</td>
<td>420</td>
<td></td>
</tr>
<tr>
<td>17.11 Exercises</td>
<td>421</td>
<td></td>
</tr>
<tr>
<td>18 Implicit Function Theorem</td>
<td>425</td>
<td></td>
</tr>
<tr>
<td>18.1 Statement And Proof Of The Theorem</td>
<td>425</td>
<td></td>
</tr>
<tr>
<td>18.2 More Derivatives</td>
<td>430</td>
<td></td>
</tr>
<tr>
<td>18.3 The Case Of $C^1$ Functions</td>
<td>431</td>
<td></td>
</tr>
<tr>
<td>18.4 Exercises</td>
<td>431</td>
<td></td>
</tr>
<tr>
<td>18.5 The Method Of Lagrange Multipliers</td>
<td>434</td>
<td></td>
</tr>
<tr>
<td>18.6 The Taylor Formula</td>
<td>435</td>
<td></td>
</tr>
<tr>
<td>18.7 Second Derivative Test</td>
<td>437</td>
<td></td>
</tr>
<tr>
<td>18.8 The Rank Theorem</td>
<td>439</td>
<td></td>
</tr>
<tr>
<td>18.9 The Local Structure Of $C^1$ Mappings</td>
<td>442</td>
<td></td>
</tr>
</tbody>
</table>
## Contents

18.10 Brouwer Fixed Point Theorem in $\mathbb{R}^n$ ................................. 444
18.10.1 Simplices and Triangulations ............................................... 445
18.10.2 Labeling Vertices ............................................................. 446
18.10.3 The Brouwer Fixed Point Theorem ...................................... 447

18.11 Invariance of Domain ............................................................. 449

18.12 Tensor Products ..................................................................... 451
18.12.1 The Norm in Tensor Product Space ...................................... 454
18.12.2 The Taylor Formula and Tensors .......................................... 456

18.13 Exercises .............................................................................. 457

19 Abstract Measures and Measurable Functions ............................... 461
19.1 Simple Functions and Measurable Functions ................................. 461
19.2 Measures and Their Properties .................................................. 465
19.3 Dynkin’s Lemma .................................................................... 466
19.4 Measures and Regularity ......................................................... 468
19.5 When Is a Measure a Borel Measure? ......................................... 471
19.6 Measures and Outer Measures ................................................. 472
19.7 Exercises .............................................................................. 473
19.8 An Outer Measure on $\mathcal{P}(\mathbb{R})$ ........................................... 475
19.9 Measures from Outer Measures .................................................. 477
19.10 One Dimensional Lebesgue-Stieltjes Measure ............................... 481
19.11 Exercises .............................................................................. 483

20 The Abstract Lebesgue Integral ..................................................... 487
20.1 Definition for Nonnegative Measurable Functions ......................... 487
20.1.1 Riemann Integrals for Decreasing Functions ......................... 487
20.2 The Lebesgue Integral for Nonnegative Simple Functions .......... 488
20.3 The Monotone Convergence Theorem ....................................... 490
20.4 Other Definitions ................................................................... 491
20.5 Fatou’s Lemma ...................................................................... 492
20.6 The Integral’s Righteous Algebraic Desires ................................. 493
20.7 The Lebesgue Integral, $L^1$ ...................................................... 494
20.8 The Dominated Convergence Theorem ..................................... 497
20.9 Exercises .............................................................................. 499

21 Measures from Positive Linear Functionals .................................... 505
21.1 Lebesgue Measure on $\mathbb{R}$; Fubini’s Theorem ............................ 511
21.2 The Besicovitch Covering Theorem .......................................... 514
21.3 Change of Variables, Linear Map .............................................. 519
21.4 Vitali Covering ...................................................................... 521
21.5 Change of Variables ................................................................ 523
21.6 Exercises .............................................................................. 529

22 The $L^p$ Spaces ........................................................................ 535
22.1 Basic Inequalities and Properties .............................................. 535
22.2 Density Considerations ............................................................ 541
22.3 Separability ........................................................................... 542
22.4 Continuity of Translation .......................................................... 544
22.5 Modifiers and Density of Smooth Functions ............................... 544
22.6 Fundamental Theorem of Calculus for Radon Measures .......... 547
22.7 Exercises .............................................................................. 550
Chapter 23: Representation Theorems

23.1 Basic Theory ............................................. 555
23.2 Radon Nikodym Theorem ................................. 559
23.3 Improved Change Of Variables Formula .............. 564
23.4 Vector Measures ........................................... 565
23.5 Representation Theorems For The Dual Space Of $L^p$ ............................................. 571
23.6 The Dual Space Of $C_0 (X)$ ................................ 577
23.7 Exercises ................................................... 582

IV Appendix ...................................................... 585

A. The Cross Product ........................................ 587
   A.1 The Box Product ......................................... 590
   A.2 The Distributive Law For Cross Product ........... 591

B. Weierstrass Approximation Theorem .................... 593
   B.1 Functions Of Many Variables ......................... 595
   B.2 Tietze Extension Theorem .............................. 600

C. The Hausdorff Maximal Theorem ......................... 603
   C.1 The Hamel Basis .......................................... 606
   C.2 Exercises ................................................ 607
Preface

This is on linear algebra and its interaction with analysis. It emphasizes the main ideas, both algebraic and geometric and attempts to present these ideas as quickly as possible without being overly terse. The emphasis will be on arbitrary fields in the first part and then later geometric ideas will be included in the context of the usual fields of $\mathbb{R}$ and $\mathbb{C}$. The first part is on linear algebra as a part of modern algebra. It avoids cluttering the presentation with geometric and analytic ideas which are really not essential to understanding these theorems. The second part is on the role of analysis in linear algebra. It is like baby functional analysis. Some analysis ideas do in fact creep in to the first part, but they are generally fairly rudimentary, occur as examples, and will have been seen in calculus. It may be that increased understanding is obtained by this kind of presentation in which that which is purely algebraic is presented first. This also involves emphasizing the minimum polynomial more than the characteristic polynomial and postponing the determinant. In each part, I have included a few related topics which are similar to ideas found in linear algebra or which have linear algebra as a fundamental part. The third part of the book involves significant ideas from analysis which depend on linear algebra.

The book is a rewritten version of an earlier book. It also includes several topics not in this other book including a chapter which is an introduction to modules and rings and much more material on analysis. However, I am not including topics from functional analysis so much. Instead, I am limiting the topics to the standard analysis involving derivatives and integrals. In fact, if everything which uses linear algebra were presented, the book would be much longer. It is limited to topics that I especially like and emphasizes finite dimensional situations.
Chapter 1

Some Prerequisite Topics

The reader should be familiar with most of the topics in this chapter. However, it is often the case that set notation is not familiar and so a short discussion of this is included first. Complex numbers are then considered in somewhat more detail. Many of the applications of linear algebra require the use of complex numbers, so this is the reason for this introduction. Then polynomials and finite fields are discussed briefly to emphasize that linear algebra works for any field of scalars, not just the field of real and complex numbers.

1.1 Sets And Set Notation

A set is just a collection of things called elements. Often these are also referred to as points in calculus. For example, \( \{1, 2, 3, 8\} \) would be a set consisting of the elements 1, 2, 3, and 8. To indicate that 3 is an element of \( \{1, 2, 3, 8\} \), it is customary to write \( 3 \in \{1, 2, 3, 8\} \). Sometimes a rule specifies a set. For example you could specify a set:

\[
\{a \in \mathbb{Z} : a > 2\}.
\]

This notation says: the set of all integers, \( a \), such that \( a > 2 \).

If \( A \) and \( B \) are sets with the property that every element of \( A \) is an element of \( B \), then \( A \) is a subset of \( B \). For example, \( \{1, 2, 3, 8\} \) is a subset of \( \{1, 2, 3, 4, 5, 8\} \), in symbols, \( \{1, 2, 3, 8\} \subseteq \{1, 2, 3, 4, 5, 8\} \). It is sometimes said that “\( A \) is contained in \( B \)” or even “\( B \) contains \( A \)”. The same statement about the two sets may also be written as \( \{1, 2, 3, 4, 5, 8\} \supseteq \{1, 2, 3, 8\} \).

The union of two sets is the set consisting of everything which is an element of at least one of the sets, \( A \) or \( B \). As an example of the union of two sets \( \{1, 2, 3, 8\} \cup \{3, 4, 7, 8\} = \{1, 2, 3, 4, 7, 8\} \) because these numbers are those which are in at least one of the two sets. In general

\[
A \cup B \equiv \{x : x \in A \text{ or } x \in B\}.
\]

Be sure you understand that something which is in both \( A \) and \( B \) is in the union. It is not an exclusive or.

The intersection of two sets, \( A \) and \( B \) consists of everything which is in both of the sets. Thus \( \{1, 2, 3, 8\} \cap \{3, 4, 7, 8\} = \{3, 8\} \) because 3 and 8 are those elements the two sets have in common. In general,

\[
A \cap B \equiv \{x : x \in A \text{ and } x \in B\}.
\]

The symbol \([a, b]\) where \( a \) and \( b \) are real numbers, denotes the set of real numbers \( x \), such that \( a \leq x \leq b \) and \([a, b]\) denotes the set of real numbers such that \( a \leq x < b \). \((a, b)\) consists of the set of real numbers \( x \) such that \( a < x < b \) and \([a, b]\) indicates the set of numbers \( x \) such that \( a < x \leq b \). \([a, \infty)\) means the set of all numbers \( x \) such that \( x \geq a \) and \((-\infty, a]\) means the set of all real numbers which are less than or equal to \( a \). These sorts of sets of real numbers are called intervals. The two points \( a \) and \( b \) are called endpoints of the interval. Other intervals such as \((-\infty, b]\) are defined by analogy to what was just explained. In general, the curved parenthesis indicates the endpoint it sits next to is not included while the square parenthesis indicates this end point is included. The
reason that there will always be a curved parenthesis next to $\infty$ or $-\infty$ is that these are not real numbers. Therefore, they cannot be included in any set of real numbers.

A special set which needs to be given a name is the empty set also called the null set, denoted by $\emptyset$. Thus $\emptyset$ is defined as the set which has no elements in it. Mathematicians like to say the empty set is a subset of every set. The reason they say this is that if it were not so, there would have to exist a set $A$, such that $\emptyset$ has something in it which is not in $A$. However, $\emptyset$ has nothing in it and so the least intellectual discomfort is achieved by saying $\emptyset \subseteq A$.

If $A$ and $B$ are two sets, $A \setminus B$ denotes the set of things which are in $A$ but not in $B$. Thus $$A \setminus B \equiv \{ x \in A : x \notin B \}.$$ Set notation is used whenever convenient.

To illustrate the use of this notation relative to intervals consider three examples of inequalities. Their solutions will be written in the notation just described.

Example 1.1.1 Solve the inequality $2x + 4 \leq x - 8$

$x \leq -12$ is the answer. This is written in terms of an interval as $(-\infty, -12]$.

Example 1.1.2 Solve the inequality $(x + 1)(2x - 3) \geq 0$.

The solution is $x \leq -1$ or $x \geq \frac{3}{2}$. In terms of set notation this is denoted by $(-\infty, -1] \cup \left[\frac{3}{2}, \infty\right)$.

Example 1.1.3 Solve the inequality $x(x + 2) \geq -4$.

This is true for any value of $x$. It is written as $\mathbb{R}$ or $(-\infty, \infty)$.

Something is in the Cartesian product of a set whose elements are sets if it consists of a single thing taken from each set in the family. Thus $(1, 2, 3) \in \{1, 4, 2\} \times \{1, 2, 7\} \times \{4, 3, 7, 9\}$ because it consists of exactly one element from each of the sets which are separated by $\times$. Also, this is the notation for the Cartesian product of finitely many sets. If $\mathcal{S}$ is a set whose elements are sets,

$$\prod_{A \in \mathcal{S}} A$$

signifies the Cartesian product.

The Cartesian product is the set of choice functions, a choice function being a function which selects exactly one element of each set of $\mathcal{S}$. You may think the axiom of choice, stating that the Cartesian product of a nonempty family of nonempty sets is nonempty, is innocuous but there was a time when many mathematicians were ready to throw it out because it implies things which are very hard to believe, things which never happen without the axiom of choice.

### 1.2 The Schroder Bernstein Theorem

It is very important to be able to compare the size of sets in a rational way. The most useful theorem in this context is the Schroder Bernstein theorem which is the main result to be presented in this section. The Cartesian product is discussed above. The next definition reviews this and defines the concept of a function.

Definition 1.2.1 Let $X$ and $Y$ be sets.

$$X \times Y \equiv \{(x, y) : x \in X \text{ and } y \in Y\}$$

A relation is defined to be a subset of $X \times Y$. A function $f$, also called a mapping, is a relation which has the property that if $(x, y)$ and $(x, y_1)$ are both elements of the $f$, then $y = y_1$. The domain of $f$ is defined as

$$D(f) \equiv \{x : (x, y) \in f\},$$
1.2. THE SCHRODER BERNSTEIN THEOREM

written as \( f : D ( f ) \to Y \). Another notation which is used is the following

\[ f^{-1} ( y ) \equiv \{ x \in D ( f ) : f ( x ) = y \} \]

This is called the inverse image.

It is probably safe to say that most people do not think of functions as a type of relation which is a subset of the Cartesian product of two sets. A function is like a machine which takes inputs, \( x \) and makes them into a unique output, \( f ( x ) \). Of course, that is what the above definition says with more precision. An ordered pair, \((x, y)\) which is an element of the function or mapping has an input, \( x \) and a unique output \( y \), denoted as \( f ( x ) \) while the name of the function is \( f \). “mapping” is often a noun meaning function. However, it also is a verb as in “\( f \) is mapping \( A \) to \( B \)”. That which a function is thought of as doing is also referred to using the word “maps” as in: \( f \) maps \( X \) to \( Y \). However, a set of functions may be called a set of maps so this word might also be used as the plural of a noun. There is no help for it. You just have to suffer with this nonsense.

The following theorem which is interesting for its own sake will be used to prove the Schroder Bernstein theorem, proved by Dedekind in 1887. The proof given here is like the version in Hewitt and Stromberg [27].

**Theorem 1.2.2** Let \( f : X \to Y \) and \( g : Y \to X \) be two functions. Then there exist sets \( A, B, C, D \), such that

\[
A \cup B = X, \quad C \cup D = Y, \quad A \cap B = \emptyset, \quad C \cap D = \emptyset, \quad f ( A ) = C, \quad g ( D ) = B.
\]

The following picture illustrates the conclusion of this theorem.

\[
\begin{array}{|c|}
\hline
X \\
\hline
A \\
\hline
B = g ( D ) \\
\hline
\end{array} \quad \begin{array}{|c|}
\hline
f \\
\hline \begin{array}{c}
C = f ( A ) \\
\end{array}
\hline
D \\
\hline
\end{array}
\]

**Proof:** Consider the empty set, \( \emptyset \subseteq X \). If \( y \in Y \setminus f ( \emptyset ) \), then \( g ( y ) \notin \emptyset \) because \( \emptyset \) has no elements. Also, if \( A, B, C, \) and \( D \) are as described above, \( A \) also would have this same property that the empty set has. However, \( A \) is probably larger. Therefore, say \( A_0 \subseteq X \) satisfies \( \mathcal{P} \) if whenever \( y \in Y \setminus f ( A_0 ) \), \( g ( y ) \notin A_0 \).

\[
A \equiv \{ A_0 \subseteq X : A_0 \text{ satisfies } \mathcal{P} \}.
\]

Let \( A = \cup A \). If \( y \in Y \setminus f ( A ) \), then for each \( A_0 \in A \), \( y \in Y \setminus f ( A_0 ) \) and so \( g ( y ) \notin A_0 \). Since \( g ( y ) \notin A_0 \) for all \( A_0 \in A \), it follows \( g ( y ) \notin A \). Hence \( A \) satisfies \( \mathcal{P} \) and is the largest subset of \( X \) which does so. Now define

\[
C \equiv f ( A ), \quad D \equiv Y \setminus C, \quad B \equiv X \setminus A.
\]

It only remains to verify that \( g ( D ) = B \). It was just shown that \( g ( D ) \subseteq B \).

Suppose \( x \in B = X \setminus A \). Then \( A \cup \{ x \} \) does not satisfy \( \mathcal{P} \) and so there exists \( y \in Y \setminus f ( A \cup \{ x \} ) \subseteq D \) such that \( g ( y ) \in A \cup \{ x \} \). But \( y \notin f ( A ) \) and so since \( A \) satisfies \( \mathcal{P} \), it follows \( g ( y ) \notin A \). Hence \( g ( y ) = x \) and so \( x \in g ( D ) \). Hence \( g ( D ) = B \). ■

**Theorem 1.2.3** (Schroder Bernstein) If \( f : X \to Y \) and \( g : Y \to X \) are one to one, then there exists \( h : X \to Y \) which is one to one and onto.
Proof: Let $A, B, C, D$ be the sets of Theorem 1.2.2 and define

$$h(x) = \begin{cases} f(x) & \text{if } x \in A \\ g^{-1}(x) & \text{if } x \in B \end{cases}$$

Then $h$ is the desired one to one and onto mapping. ■

Recall that the Cartesian product may be considered as the collection of choice functions.

Definition 1.2.4 Let $I$ be a set and let $X_i$ be a set for each $i \in I$. $f$ is a choice function written as

$$f \in \prod_{i \in I} X_i$$

if $f(i) \in X_i$ for each $i \in I$.

The axiom of choice says that if $X_i \neq \emptyset$ for each $i \in I$, for $I$ a set, then

$$\prod_{i \in I} X_i \neq \emptyset.$$ 

Sometimes the two functions, $f$ and $g$ are onto but not one to one. It turns out that with the axiom of choice, a similar conclusion to the above may be obtained.

Corollary 1.2.5 If $f : X \rightarrow Y$ is onto and $g : Y \rightarrow X$ is onto, then there exists $h : X \rightarrow Y$ which is one to one and onto.

Proof: For each $y \in Y$, $f^{-1}(y) \equiv \{ x \in X : f(x) = y \} \neq \emptyset$. Therefore, by the axiom of choice, there exists $f_0^{-1} \in \prod_{y \in Y} f^{-1}(y)$ which is the same as saying that for each $y \in Y$, $f_0^{-1}(y) \in f^{-1}(y)$. Similarly, there exists $g_0^{-1}(x) \in g^{-1}(x)$ for all $x \in X$. Then $f_0^{-1}$ is one to one because if $f_0^{-1}(y_1) = f_0^{-1}(y_2)$, then

$$y_1 = f(f_0^{-1}(y_1)) = f(f_0^{-1}(y_2)) = y_2.$$ 

Similarly $g_0^{-1}$ is one to one. Therefore, by the Schroder Bernstein theorem, there exists $h : X \rightarrow Y$ which is one to one and onto. ■

Definition 1.2.6 A set $S$, is finite if there exists a natural number $n$ and a map $\theta$ which maps $\{1, \cdots , n\}$ one to one and onto $S$. $S$ is infinite if it is not finite. A set $S$, is called countable if there exists a map $\theta$ mapping $\mathbb{N}$ one to one and onto $S$. (When $\theta$ maps a set $A$ to a set $B$, this will be written as $\theta : A \rightarrow B$ in the future.) Here $\mathbb{N} \equiv \{1, 2, \cdots \}$, the natural numbers. $S$ is at most countable if there exists a map $\theta : \mathbb{N} \rightarrow S$ which is onto.

The property of being at most countable is often referred to as being countable because the question of interest is normally whether one can list all elements of the set, designating a first, second, third etc. in such a way as to give each element of the set a natural number. The possibility that a single element of the set may be counted more than once is often not important.

Theorem 1.2.7 If $X$ and $Y$ are both at most countable, then $X \times Y$ is also at most countable. If either $X$ or $Y$ is countable, then $X \times Y$ is also countable.

Proof: It is given that there exists a mapping $\eta : \mathbb{N} \rightarrow X$ which is onto. Define $\eta(i) \equiv x_i$ and consider $X$ as the set $\{x_1, x_2, x_3, \cdots \}$. Similarly, consider $Y$ as the set $\{y_1, y_2, y_3, \cdots \}$. It follows the elements of $X \times Y$ are included in the following rectangular array.

$$\begin{array}{cccc} (x_1, y_1) & (x_1, y_2) & (x_1, y_3) & \cdots \\ (x_2, y_1) & (x_2, y_2) & (x_2, y_3) & \cdots \\ (x_3, y_1) & (x_3, y_2) & (x_3, y_3) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{array}$$
Follow a path through this array as follows.

\[
\begin{array}{c}
(x_1, y_1) \rightarrow (x_1, y_2) \rightarrow (x_1, y_3) \\
(x_2, y_1) \rightarrow (x_2, y_2) \\
\downarrow \\
(x_3, y_1)
\end{array}
\]

Thus the first element of \(X \times Y\) is \((x_1, y_1)\), the second element of \(X \times Y\) is \((x_1, y_2)\), the third element of \(X \times Y\) is \((x_2, y_1)\) etc. This assigns a number from \(\mathbb{N}\) to each element of \(X \times Y\). Thus \(X \times Y\) is at most countable.

It remains to show the last claim. Suppose without loss of generality that \(X\) is countable. Then there exists \(\alpha : \mathbb{N} \rightarrow X\) which is one to one and onto. Let \(\beta : X \times Y \rightarrow \mathbb{N}\) be defined by \(\beta ((x, y)) = \alpha^{-1}(x)\). Thus \(\beta\) is onto \(\mathbb{N}\). By the first part there exists a function from \(\mathbb{N}\) onto \(X \times Y\). Therefore, by Corollary 1.2.5, there exists a one to one and onto mapping from \(X \times Y\) to \(\mathbb{N}\). □

**Theorem 1.2.8** If \(X\) and \(Y\) are at most countable, then \(X \cup Y\) is at most countable. If either \(X\) or \(Y\) are countable, then \(X \cup Y\) is countable.

**Proof:** As in the preceding theorem,

\[
X = \{x_1, x_2, x_3, \cdots\}
\]

and

\[
Y = \{y_1, y_2, y_3, \cdots\}.
\]

Consider the following array consisting of \(X \cup Y\) and path through it.

\[
\begin{array}{c}
x_1 \rightarrow x_2 \rightarrow x_3 \\
y_1 \rightarrow y_2
\end{array}
\]

Thus the first element of \(X \cup Y\) is \(x_1\), the second is \(x_2\) the third is \(y_1\) the fourth is \(y_2\) etc.

Consider the second claim. By the first part, there is a map from \(\mathbb{N}\) onto \(X \times Y\). Suppose without loss of generality that \(X\) is countable and \(\alpha : \mathbb{N} \rightarrow X\) is one to one and onto. Then define \(\beta (y) \equiv 1\), for all \(y \in Y\) and \(\beta (x) \equiv \alpha^{-1}(x)\). Thus, \(\beta\) maps \(X \times Y\) onto \(\mathbb{N}\) and this shows there exist two onto maps, one mapping \(X \cup Y\) onto \(\mathbb{N}\) and the other mapping \(\mathbb{N}\) onto \(X \cup Y\). Then Corollary 1.2.5 yields the conclusion. □

Note that by induction this shows that if you have any finite set whose elements are countable sets, then the union of these is countable.

### 1.3 Equivalence Relations

There are many ways to compare elements of a set other than to say two elements are equal or the same. For example, in the set of people let two people be equivalent if they have the same weight. This would not be saying they were the same person, just that they weighed the same. Often such relations involve considering one characteristic of the elements of a set and then saying the two elements are equivalent if they are the same as far as the given characteristic is concerned.

**Definition 1.3.1** Let \(S\) be a set. \(\sim\) is an equivalence relation on \(S\) if it satisfies the following axioms.

1. \(x \sim x\) for all \(x \in S\). (Reflexive)
2. If $x \sim y$ then $y \sim x$. (Symmetric)

3. If $x \sim y$ and $y \sim z$, then $x \sim z$. (Transitive)

**Definition 1.3.2** $[x]$ denotes the set of all elements of $S$ which are equivalent to $x$ and $[x]$ is called the equivalence class determined by $x$ or just the equivalence class of $x$.

With the above definition one can prove the following simple theorem.

**Theorem 1.3.3** Let $\sim$ be an equivalence class defined on a set, $S$ and let $\mathcal{H}$ denote the set of equivalence classes. Then if $[x]$ and $[y]$ are two of these equivalence classes, either $x \sim y$ and $[x] = [y]$ or it is not true that $x \sim y$ and $[x] \cap [y] = \emptyset$.

### 1.4 Well Ordering And Induction

Mathematical induction and well ordering are two extremely important principles in math. They are often used to prove significant things which would be hard to prove otherwise.

**Definition 1.4.1** A set is well ordered if every nonempty subset $S$ contains a smallest element $z$ having the property that $z \leq x$ for all $x \in S$.

**Axiom 1.4.2** Any set of integers larger than a given number is well ordered.

In particular, the natural numbers defined as

$$\mathbb{N} \equiv \{1, 2, \cdots\}$$

is well ordered.

The above axiom implies the principle of mathematical induction. The symbol $\mathbb{Z}$ denotes the set of all integers. Note that if $a$ is an integer, then there are no integers between $a$ and $a + 1$.

**Theorem 1.4.3** (Mathematical induction) A set $S \subseteq \mathbb{Z}$, having the property that $a \in S$ and $n+1 \in S$ whenever $n \in S$ contains all integers $x \in \mathbb{Z}$ such that $x \geq a$.

**Proof:** Let $T$ consist of all integers larger than or equal to $a$ which are not in $S$. The theorem will be proved if $T = \emptyset$. If $T \neq \emptyset$ then by the well ordering principle, there would have to exist a smallest element of $T$, denoted as $b$. It must be the case that $b > a$ since by definition, $a \notin T$. Thus $b \geq a + 1$, and so $b - 1 \geq a$ and $b - 1 \notin S$ because if $b - 1 \in S$, then $b - 1 + 1 = b \in S$ by the assumed property of $S$. Therefore, $b - 1 \in T$ which contradicts the choice of $b$ as the smallest element of $T$. ($b - 1$ is smaller.) Since a contradiction is obtained by assuming $T \neq \emptyset$, it must be the case that $T = \emptyset$ and this says that every integer at least as large as $a$ is also in $S$.

Mathematical induction is a very useful device for proving theorems about the integers.

**Example 1.4.4** Prove by induction that $\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$.

By inspection, if $n = 1$ then the formula is true. The sum yields 1 and so does the formula on the right. Suppose this formula is valid for some $n \geq 1$ where $n$ is an integer. Then

$$\sum_{k=1}^{n+1} k^2 = \sum_{k=1}^{n} k^2 + (n + 1)^2 = \frac{n(n+1)(2n+1)}{6} + (n + 1)^2.$$

The step going from the first to the second line is based on the assumption that the formula is true for $n$. This is called the induction hypothesis. Now simplify the expression in the second line,

$$\frac{n(n+1)(2n+1)}{6} + (n + 1)^2.$$
This equals
\[(n+1)\left(\frac{n(2n+1)}{6} + (n+1)\right)\]

and
\[\frac{n(2n+1)}{6} + (n+1) = \frac{6(n+1) + 2n^2 + n}{6} = \frac{(n+2)(2n+3)}{6}\]

Therefore,
\[\sum_{k=1}^{n+1} k^2 = \frac{(n+1)(n+2)(2n+3)}{6} = \frac{(n+1)((n+1)+1)(2(n+1)+1)}{6},\]

showing the formula holds for \(n+1\) whenever it holds for \(n\). This proves the formula by mathematical induction.

**Example 1.4.5** Show that for all \(n \in \mathbb{N}\), \(\frac{1}{2} \cdot \frac{3}{4} \cdot \frac{5}{6} \cdots \frac{2n-1}{2n} < \frac{1}{\sqrt{2n+1}}\).

If \(n = 1\) this reduces to the statement that \(\frac{1}{2} < \frac{1}{\sqrt{3}}\), which is obviously true. Suppose then that the inequality holds for \(n\). Then
\[\frac{1}{2} \cdot \frac{3}{4} \cdot \frac{5}{6} \cdots \frac{2n-1}{2n} \cdot \frac{2n+1}{2n+2} < \frac{1}{\sqrt{2n+1}} \cdot \frac{2n+1}{2n+2} = \frac{\sqrt{2n+1}}{2n+2}.\]

The theorem will be proved if this last expression is less than \(\frac{1}{\sqrt{2n+3}}\). This happens if and only if
\[\left(\frac{1}{\sqrt{2n+3}}\right)^2 = \frac{1}{2n+3} > \frac{2n+1}{(2n+2)^2}\]

which occurs if and only if \((2n+2)^2 > (2n+3)(2n+1)\) and this is clearly true which may be seen from expanding both sides. This proves the inequality.

Let’s review the process just used. If \(S\) is the set of integers at least as large as 1 for which the formula holds, the first step was to show \(1 \in S\) and then that whenever \(n \in S\), it follows \(n+1 \in S\). Therefore, by the principle of mathematical induction, \(S\) contains \([1, \infty) \cap \mathbb{Z}\), all positive integers. In doing an inductive proof of this sort, the set \(S\) is normally not mentioned. One just verifies the steps above. First show the thing is true for some \(a \in \mathbb{Z}\) and then verify that whenever it is true for \(m\) it follows it is also true for \(m+1\). When this has been done, the theorem has been proved for all \(m \geq a\).

### 1.5 The Complex Numbers And Fields

Recall that a real number is a point on the real number line. Just as a real number should be considered as a point on the line, a complex number is considered a point in the plane which can be identified in the usual way using the Cartesian coordinates of the point. Thus \((a, b)\) identifies a point whose \(x\) coordinate is \(a\) and whose \(y\) coordinate is \(b\). In dealing with complex numbers, such a point is written as \(a + bi\). For example, in the following picture, I have graphed the point \(3 + 2i\). You see it corresponds to the point in the plane whose coordinates are \((3, 2)\).

Multiplication and addition are defined in the most obvious way subject to the convention that \(i^2 = -1\). Thus,
\[(a + ib) + (c + id) = (a + c) + i(b + d)\]

and
\[(a + ib)(c + id) = ac + iad + ibc + i^2bd = (ac - bd) + i(bc + ad)\].
Every non zero complex number \( a + ib \), with \( a^2 + b^2 \neq 0 \), has a unique multiplicative inverse.

\[
\frac{1}{a + ib} = \frac{a - ib}{a^2 + b^2} = \frac{a}{a^2 + b^2} - \frac{ib}{a^2 + b^2}.
\]

You should prove the following theorem.

**Theorem 1.5.1** The complex numbers with multiplication and addition defined as above form a field satisfying all the field axioms. These are the following list of properties. In this list, \( \mathbb{F} \) is the symbol for a field.

1. \( x + y = y + x, \) (commutative law for addition)
2. There exists \( 0 \) such that \( x + 0 = x \) for all \( x, \) (additive identity).
3. For each \( x \in \mathbb{F}, \) there exists \( -x \in \mathbb{F} \) such that \( x + (-x) = 0, \) (existence of additive inverse).
4. \( (x + y) + z = x + (y + z), \) (associative law for addition).
5. \( xy = yx, \) (commutative law for multiplication). You could write this as \( x \times y = y \times x. \)
6. \( (xy)z = x(yz), \) (associative law for multiplication).
7. There exists \( 1 \) such that \( 1x = x \) for all \( x, \) (multiplicative identity).
8. For each \( x \neq 0, \) there exists \( x^{-1} \) such that \( xx^{-1} = 1, \) (existence of multiplicative inverse).
9. \( x(y + z) = xy + xz, \) (distributive law).

When you have a field \( \mathbb{F} \) some things follow right away from the above axioms.

**Theorem 1.5.2** Let \( \mathbb{F} \) be a field. This means it satisfies the axioms of the above theorem. Then the following hold.

1. \( 0 \) is unique.
2. \( -x \) is unique
3. \( 0x = 0 \)
4. \( (-1)x = -x \)
5. \( x^{-1} \) is unique

**Proof:** Consider the first claim. Suppose \( \hat{0} \) is another additive identity. Then

\[
\hat{0} = \hat{0} + 0 = 0
\]

and so sure enough, there is only one such additive identity. Consider uniqueness of \( -x \) next. Suppose \( y \) is also an additive inverse. Then

\[
-x = -x + 0 = -x + (x + y) = (-x + x) + y = 0 + y = y
\]

so the additive inverse is unique also.

\[
0x = (0 + 0)x = 0x + 0x
\]

Now add \(-0x\) to both sides to conclude that \( 0 = 0x. \) Next

\[
0 = (1 + (-1))x = x + (-1)x
\]
and by uniqueness of \(-x\), this implies \((-1) \cdot x = -x\) as claimed. Finally, if \(x \neq 0\) and \(y\) is a multiplicative inverse,

\[
x^{-1} = 1x^{-1} = (yx) x^{-1} = y(xx^{-1}) = y1 = y
\]

so \(y = x^{-1}\).

Something which satisfies these axioms is called a field. Linear algebra is all about fields, although in this book, the field of most interest will be the field of complex numbers or the field of real numbers. You have seen in earlier courses that the real numbers also satisfy the above axioms. The field of complex numbers is denoted as \(C\) and the field of real numbers is denoted as \(R\). An important construction regarding complex numbers is the complex conjugate denoted by a horizontal line above the number. It is defined as follows.

\[
a + ib \equiv a - ib.
\]

What it does is reflect a given complex number across the \(x\) axis. Algebraically, the following formula is easy to obtain.

\[
(a + ib)(a + ib) = (a - ib)(a + ib) = a^2 + b^2 - i(ab - ab) = a^2 + b^2.
\]

**Definition 1.5.3** Define the absolute value of a complex number as follows.

\[
|a + ib| \equiv \sqrt{a^2 + b^2}.
\]

Thus, denoting by \(z\) the complex number \(z = a + ib\),

\[
|z| = (z\bar{z})^{1/2}.
\]

Also from the definition, if \(z = x + iy\) and \(w = u + iv\) are two complex numbers, then \(|zw| = |z||w|\). You should verify this.

**Notation 1.5.4** Recall the following notation.

\[
\sum_{j=1}^{n} a_j \equiv a_1 + \cdots + a_n
\]

There is also a notation which is used to denote a product.

\[
\prod_{j=1}^{n} a_j \equiv a_1 a_2 \cdots a_n
\]

The triangle inequality holds for the absolute value for complex numbers just as it does for the ordinary absolute value.

**Proposition 1.5.5** Let \(z, w\) be complex numbers. Then the triangle inequality holds.

\[
|z + w| \leq |z| + |w|, \quad ||z| - |w|| \leq |z - w|.
\]

**Proof:** Let \(z = x + iy\) and \(w = u + iv\). First note that

\[
z\bar{w} = (x + iy)(u - iv) = xu + yv + i(yu - xv)
\]

and so \(|xu + yv| \leq |z\bar{w}| = |z||w|\).

\[
|z + w|^2 = (x + u + i(y + v))(x + u - i(y + v))
\]

\[
= (x + u)^2 + (y + v)^2 = x^2 + u^2 + 2xu + 2yv + y^2 + v^2.
\]
10

CHAPTER 1. SOME PREREQUISITE TOPICS

\[ \leq |z|^2 + |w|^2 + 2|z||w| = (|z| + |w|)^2, \]

so this shows the first version of the triangle inequality. To get the second,

\[ z = z - w + w, \quad w = w - z + z \]

and so by the first form of the inequality

\[ |z| \leq |z - w| + |w|, \quad |w| \leq |z - w| + |z| \]

and so both \(|z| - |w|\) and \(|w| - |z|\) are no larger than \(|z - w|\) and this proves the second version because \(||z| - |w||\) is one of \(|z| - |w|\) or \(|w| - |z|\). \[ \square \]

With this definition, it is important to note the following. Be sure to verify this. It is not too hard but you need to do it.

**Remark 1.5.6**: Let \(z = a + ib\) and \(w = c + id\). Then \(|z - w| = \sqrt{(a - c)^2 + (b - d)^2}\). Thus the distance between the point in the plane determined by the ordered pair \((a, b)\) and the ordered pair \((c, d)\) equals \(|z - w|\) where \(z\) and \(w\) are as just described.

For example, consider the distance between \((2, 5)\) and \((1, 8)\). From the distance formula this distance equals \(\sqrt{(2 - 1)^2 + (5 - 8)^2} = \sqrt{10}\). On the other hand, letting \(z = 2 + i5\) and \(w = 1 + i8\), \(z - w = 1 - i3\) and so \((z - w)(\overline{z - w}) = (1 - i3)(1 + i3) = 10\) so \(|z - w| = \sqrt{10}\), the same thing obtained with the distance formula.

### 1.6 Polar Form Of Complex Numbers

Complex numbers, are often written in the so called polar form which is described next. Suppose \(z = x + iy\) is a complex number. Then

\[ x + iy = \sqrt{x^2 + y^2} \left( \frac{x}{\sqrt{x^2 + y^2}} + \frac{y}{\sqrt{x^2 + y^2}} i \right). \]

Now note that

\[ \left( \frac{x}{\sqrt{x^2 + y^2}} \right)^2 + \left( \frac{y}{\sqrt{x^2 + y^2}} \right)^2 = 1 \]

and so

\[ \left( \frac{x}{\sqrt{x^2 + y^2}} \right) - \left( \frac{y}{\sqrt{x^2 + y^2}} \right) \]

is a point on the unit circle. Therefore, there exists a unique angle \(\theta \in [0, 2\pi)\) such that

\[ \cos \theta = \frac{x}{\sqrt{x^2 + y^2}}, \quad \sin \theta = \frac{y}{\sqrt{x^2 + y^2}}. \]

The polar form of the complex number is then \(r (\cos \theta + i \sin \theta)\) where \(\theta\) is this angle just described and \(r = \sqrt{x^2 + y^2} \equiv |z|\).
1.7 Roots Of Complex Numbers

A fundamental identity is the formula of De Moivre which follows.

**Theorem 1.7.1** Let \( r > 0 \) be given. Then if \( n \) is a positive integer,
\[
[r (\cos t + i \sin t)]^n = r^n (\cos nt + i \sin nt).
\]

**Proof:** It is clear the formula holds if \( n = 1 \). Suppose it is true for \( n \).
\[
[r (\cos t + i \sin t)]^{n+1} = [r (\cos t + i \sin t)]^n [r (\cos t + i \sin t)]
\]
which by induction equals
\[
r^{n+1} (\cos nt + i \sin nt) (\cos t + i \sin t)
\]
\[
= r^{n+1} ((\cos nt \cos t - \sin nt \sin t) + i (\sin nt \cos t + \cos nt \sin t))
\]
\[
= r^{n+1} (\cos (n+1)t + i \sin (n+1)t)
\]
by the formulas for the cosine and sine of the sum of two angles. \( \blacksquare \)

**Corollary 1.7.2** Let \( z \) be a non zero complex number. Then there are always exactly \( k \) \( k^{th} \) roots of \( z \) in \( \mathbb{C} \).

**Proof:** Let \( z = x + iy \) and let \( z = |z| (\cos t + i \sin t) \) be the polar form of the complex number. By De Moivre’s theorem, a complex number
\[
r (\cos \alpha + i \sin \alpha),
\]
is a \( k^{th} \) root of \( z \) if and only if
\[
r^k (\cos k\alpha + i \sin k\alpha) = |z| (\cos t + i \sin t).
\]
This requires \( r^k = |z| \) and so \( r = |z|^{1/k} \) and also both \( \cos (k\alpha) = \cos t \) and \( \sin (k\alpha) = \sin t \). This can only happen if
\[
k\alpha = t + 2l\pi
\]
for \( l \) an integer. Thus
\[
\alpha = \frac{t + 2l\pi}{k}, \ l \in \mathbb{Z}
\]
and so the \( k^{th} \) roots of \( z \) are of the form
\[
|z|^{1/k}\left( \cos \left( \frac{t + 2l\pi}{k} \right) + i \sin \left( \frac{t + 2l\pi}{k} \right) \right), \ l \in \mathbb{Z}.
\]
Since the cosine and sine are periodic of period \( 2\pi \), there are exactly \( k \) distinct numbers which result from this formula. \( \blacksquare \)

**Example 1.7.3** Find the three cube roots of \( i \).

First note that \( i = 1 (\cos \left( \frac{\pi}{2} \right) + i \sin \left( \frac{\pi}{2} \right)) \). Using the formula in the proof of the above corollary, the cube roots of \( i \) are
\[
1 \left( \cos \left( \frac{\pi/2 + 2l\pi}{3} \right) + i \sin \left( \frac{\pi/2 + 2l\pi}{3} \right) \right)
\]
where \( l = 0, 1, 2 \). Therefore, the roots are
\[
\cos \left( \frac{\pi}{6} \right) + i \sin \left( \frac{\pi}{6} \right), \cos \left( \frac{5\pi}{6} \right) + i \sin \left( \frac{5\pi}{6} \right), \cos \left( \frac{3\pi}{2} \right) + i \sin \left( \frac{3\pi}{2} \right).
\]
Thus the cube roots of \( i \) are \( \frac{\sqrt{3}}{2} + i \left( \frac{1}{2} \right), -\frac{\sqrt{3}}{2} + i \left( \frac{1}{2} \right), \text{ and } -i \).

The ability to find \( k^{th} \) roots can also be used to factor some polynomials.
Example 1.7.4 Factor the polynomial $x^3 - 27$.

First find the cube roots of $27$. By the above procedure using De Moivre’s theorem, these cube roots are $3\left(\frac{-1}{2} + \frac{i\sqrt{3}}{2}\right)$, and $3\left(\frac{-1}{2} - \frac{i\sqrt{3}}{2}\right)$. Therefore, $x^3 - 27 =$

$$(x - 3) \left( x - 3 \left(\frac{-1}{2} + \frac{i\sqrt{3}}{2}\right) \right) \left( x - 3 \left(\frac{-1}{2} - \frac{i\sqrt{3}}{2}\right) \right).$$

Note also $(x - 3 \left(\frac{-1}{2} + \frac{i\sqrt{3}}{2}\right)) \left( x - 3 \left(\frac{-1}{2} - \frac{i\sqrt{3}}{2}\right) \right) = x^2 + 3x + 9$ and so

$$x^3 - 27 = (x - 3) \left( x^2 + 3x + 9 \right)$$

where the quadratic polynomial $x^2 + 3x + 9$ cannot be factored without using complex numbers.

Note that even though the polynomial $x^3 - 27$ has all real coefficients, it has some complex zeros, $\frac{-1}{2} + \frac{i\sqrt{3}}{2}$ and $\frac{-1}{2} - \frac{i\sqrt{3}}{2}$. These zeros are complex conjugates of each other. It is always this way. You should show this is the case. To see how to do this, see Problems 17 and 18 below.

Another fact for your information is the fundamental theorem of algebra. This theorem says that any polynomial of degree at least 1 having any complex coefficients always has a root in $\mathbb{C}$. This is sometimes referred to by saying $\mathbb{C}$ is algebraically complete. Gauss is usually credited with giving a proof of this theorem in 1797 but many others worked on it and the first completely correct proof was due to Argand in 1806. For more on this theorem, you can google fundamental theorem of algebra and look at the interesting Wikipedia article on it. Proofs of this theorem usually involve the use of techniques from calculus even though it is really a result in algebra. A proof and plausibility explanation is given later.

1.8 The Quadratic Formula

The quadratic formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

gives the solutions $x$ to

$$ax^2 + bx + c = 0$$

where $a, b, c$ are real numbers. It holds even if $b^2 - 4ac < 0$. This is easy to show from the above. There are exactly two square roots to this number $b^2 - 4ac$ from the above methods using De Moivre’s theorem. These roots are of the form

$$\sqrt{4ac - b^2} \left( \cos \left( \frac{\pi}{2} \right) + i \sin \left( \frac{\pi}{2} \right) \right) = i\sqrt{4ac - b^2}$$

and

$$\sqrt{4ac - b^2} \left( \cos \left( \frac{3\pi}{2} \right) + i \sin \left( \frac{3\pi}{2} \right) \right) = -i\sqrt{4ac - b^2}$$

Thus the solutions, according to the quadratic formula are still given correctly by the above formula.

Do these solutions predicted by the quadratic formula continue to solve the quadratic equation? Yes, they do. You only need to observe that when you square a square root of a complex number $z$, you recover $z$. Thus

$$a \left( \frac{-b + \sqrt{b^2 - 4ac}}{2a} \right)^2 + b \left( \frac{-b + \sqrt{b^2 - 4ac}}{2a} \right) + c$$

$$= a \left( \frac{1}{2a^2} b^2 - \frac{1}{a} c - \frac{1}{2a^2} b\sqrt{b^2 - 4ac} \right) + b \left( \frac{-b + \sqrt{b^2 - 4ac}}{2a} \right) + c$$
1.9. THE COMPLEX EXPONENTIAL

\[= \left( \frac{1}{2a} \left( b\sqrt{b^2 - 4ac + 2ac - b^2} \right) \right) + \frac{1}{2a} \left( b\sqrt{b^2 - 4ac - b^2} \right) + c = 0 \]

Similar reasoning shows directly that \(\frac{-b - \sqrt{b^2 - 4ac}}{2a} \) also solves the quadratic equation.

What if the coefficients of the quadratic equation are actually complex numbers? Does the formula hold even in this case? The answer is yes. This is a hint on how to do Problem 27 below, a special case of the fundamental theorem of algebra, and an ingredient in the proof of some versions of this theorem.

Example 1.8.1 Find the solutions to \(x^2 - 2ix - 5 = 0\).

Formally, from the quadratic formula, these solutions are

\[x = \frac{2i \pm \sqrt{-4 + 20}}{2} = \frac{2i \pm 4}{2} = i \pm 2.\]

Now you can check that these really do solve the equation. In general, this will be the case. See Problem 27 below.

1.9 The Complex Exponential

It was shown above that every complex number can be written in the form \(r (\cos \theta + i \sin \theta)\) where \(r \geq 0\). Laying aside the zero complex number, this shows that every non zero complex number is of the form \(e^{i \beta} \). We write this in the form \(e^{\alpha + i \beta}\). Having done so, does it follow that the expression preserves the most important property of the function \(t \mapsto e^{(\alpha + i \beta)t}\) for \(t\) real, that

\[\left( e^{(\alpha + i \beta)t} \right)' = (\alpha + i \beta) e^{(\alpha + i \beta)t}.\]

By the definition just given which does not contradict the usual definition in case \(\beta = 0\) and the usual rules of differentiation in calculus,

\[\left( e^{(\alpha + i \beta)t} \right)' = (e^{\alpha t} (\cos (\beta t) + i \sin (\beta t)))' = e^{\alpha t} [\alpha (\cos (\beta t) + i \sin (\beta t)) + (\beta \sin (\beta t) + i \beta \cos (\beta t))]\]

Now consider the other side. From the definition it equals

\[(\alpha + i \beta) \left( e^{\alpha t} (\cos (\beta t) + i \sin (\beta t)) \right) = e^{\alpha t} [\alpha (\cos (\beta t) + i \sin (\beta t))] = e^{\alpha t} \left[ (\cos (\beta t) + i \sin (\beta t)) + (\beta \sin (\beta t) + i \beta \cos (\beta t)) \right]\]

which is the same thing. This is of fundamental importance in differential equations. It shows that there is no change in going from real to complex numbers for \(\omega\) in the consideration of the problem \(y' = \omega y, \ y(0) = 1\). The solution is always \(e^{\omega t}\). The formula just discussed, that

\[e^{\alpha (\cos \beta + i \sin \beta)} = e^{\alpha + i \beta}\]

is Euler’s formula.

1.10 The Fundamental Theorem Of Algebra

The fundamental theorem of algebra states that every non constant polynomial having coefficients in \(\mathbb{C}\) has a zero in \(\mathbb{C}\). If \(\mathbb{C}\) is replaced by \(\mathbb{R}\), this is not true because of the example, \(x^2 + 1 = 0\). This theorem is a very remarkable result and notwithstanding its title, all the most straightforward proofs depend on either analysis or topology. It was first mostly proved by Gauss in 1797. The first complete proof was given by Argand in 1806. The proof given here follows Rudin \[28\]. See also Hardy \[15\] for a similar proof, more discussion and references. The shortest proof is found in the theory of complex analysis. First I will give an informal explanation of this theorem which shows why it is reasonable to believe in the fundamental theorem of algebra.
Theorem 1.10.1 Let \( p(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 \) where each \( a_k \) is a complex number and \( a_n \neq 0, n \geq 1 \). Then there exists \( w \in \mathbb{C} \) such that \( p(w) = 0 \).

To begin with, here is the informal explanation. Dividing by the leading coefficient \( a_n \), there is no loss of generality in assuming that the polynomial is of the form

\[
p(z) = z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0
\]

If \( a_0 = 0 \), there is nothing to prove because \( p(0) = 0 \). Therefore, assume \( a_0 \neq 0 \). From the polar form of a complex number \( z \), it can be written as \( |z| (\cos \theta + i \sin \theta) \). Thus, by De Moivre’s theorem,

\[
z^n = |z|^n (\cos (n\theta) + i \sin (n\theta))
\]

It follows that \( z^n \) is some point on the circle of radius \( |z|^n \).

Denote by \( C_r \) the circle of radius \( r \) in the complex plane which is centered at 0. Then if \( r \) is sufficiently large and \( |z| = r \), the term \( z^n \) is far larger than the rest of the polynomial. It is on the circle of radius \( |z|^n \) while the other terms are on circles of fixed multiples of \( |z|^k \) for \( k \leq n - 1 \). Thus, for \( r \) large enough, \( A_r = \{ p(z) : z \in C_r \} \) describes a closed curve which misses the inside of some circle having 0 as its center. It won’t be as simple as suggested in the following picture, but it will be a closed curve thanks to De Moivre’s theorem and the observation that the cosine and sine are periodic.

Now shrink \( r \). Eventually, for \( r \) small enough, the non constant terms are negligible and so \( A_r \) is a curve which is contained in some circle centered at \( a_0 \) which has 0 on the outside.

Thus it is reasonable to believe that for some \( r \) during this shrinking process, the set \( A_r \) must hit 0. It follows that \( p(z) = 0 \) for some \( z \).

For example, consider the polynomial \( x^3 + x + 1 + i \). It has no real zeros. However, you could let \( z = r (\cos t + i \sin t) \) and insert this into the polynomial. Thus you would want to find a point where

\[
(r (\cos t + i \sin t))^3 + r (\cos t + i \sin t) + 1 + i = 0 + 0i
\]

Expanding this expression on the left to write it in terms of real and imaginary parts, you get on the left

\[
r^3 \cos^3 t - 3r^3 \cos t \sin^2 t + r \cos t + 1 + i \left( 3r^3 \cos^2 t \sin t - r^3 \sin^3 t + r \sin t + 1 \right)
\]

Thus you need to have both the real and imaginary parts equal to 0. In other words, you need to have \( (0,0) =
\[
(r^3 \cos^3 t - 3r^3 \cos t \sin^2 t + r \cos t + 1, 3r^3 \cos^2 t \sin t - r^3 \sin^3 t + r \sin t + 1)
\]

for some value of \( r \) and \( t \). First here is a graph of this parametric function of \( t \) for \( t \in [0, 2\pi] \) on the left, when \( r = 4 \). Note how the graph misses the origin \( 0 + i0 \). In fact, the closed curve is in the exterior of a circle which has the point \( 0 + i0 \) on its inside.

Next is the graph when \( r = .5 \). Note how the closed curve is included in a circle which has \( 0 + i0 \) on its outside. As you shrink \( r \) you get closed curves. At first, these closed curves enclose \( 0 + i0 \) and later, they exclude \( 0 + i0 \). Thus one of them should pass through this point. In fact, consider the curve which results when \( r = 1.386 \) which is the graph on the right. Note how for this value of \( r \)
1.11. ORDERED FIELDS

To do linear algebra, you need a field which is something satisfying the axioms listed in Theorem 1.11.1. This is generally all that is needed to do linear algebra but for the sake of completeness, the concept of an ordered field is considered here. The real numbers also have an order defined on them. This order may be defined by reference to the positive real numbers, those to the right of 0 on the number line, denoted by $\mathbb{R}^+$. More generally, for a field, one could consider an order if there is such a “positive cone” called the positive numbers such that the following axioms hold.

**Axiom 1.11.1** The sum of two positive real numbers is positive.

**Axiom 1.11.2** The product of two positive real numbers is positive.

**Axiom 1.11.3** For a given real number $x$ one and only one of the following alternatives holds. Either $x$ is positive, $x = 0$, or $-x$ is positive.

An example of this is the field of rational numbers.

**Definition 1.11.4** $x < y$ exactly when $y + (-x) \equiv y - x \in \mathbb{R}^+$. In the usual way, $x < y$ is the same as $y > x$ and $x \leq y$ means either $x < y$ or $x = y$. The symbol $\geq$ is defined similarly.

**Theorem 1.11.5** The following hold for the order defined as above.

1. If $x < y$ and $y < z$ then $x < z$ (Transitive law).
2. If $x < y$ then $x + z < y + z$ (addition to an inequality).
3. If $x \leq 0$ and $y \leq 0$, then $xy \geq 0$.
4. If $x > 0$ then $x^{-1} > 0$.
5. If $x < 0$ then $x^{-1} < 0$.
6. If $x < y$ then $xz < yz$ if $z > 0$, (multiplication of an inequality).
7. If $x < y$ and $z < 0$, then $xz > yz$ (multiplication of an inequality).
8. Each of the above holds with $>$ and $<$ replaced by $\geq$ and $\leq$ respectively except for 6 and 7 in which we must also stipulate that $x \neq 0$. 

the curve passes through the point $0 + i0$. Thus for some $t$, $1.386 (\cos t + i \sin t)$ is a solution of the equation $p(z) = 0$ or very close to one.

Now here is a short rigorous proof for those who have studied analysis. The needed analysis will be presented later in the book. You need the extreme value theorem for example.

**Proof:** Suppose the nonconstant polynomial $p(z) = a_0 + a_1 z + \cdots + a_n z^n, a_n \neq 0$, has no zero in $\mathbb{C}$. Since $\lim_{|z| \to \infty} |p(z)| = \infty$, there is a $z_0$ with $|p(z_0)| = \min_{z \in \mathbb{C}} |p(z)| > 0$

Then let $q(z) = \frac{p(z + z_0)}{p(z_0)}$. This is also a polynomial which has no zeros and the minimum of $|q(z)|$ is 1 and occurs at $z = 0$. Since $q(0) = 1$, it follows $q(z) = 1 + a_k z^k + r(z)$ where $r(z)$ consists of higher order terms. Here $a_k$ is the first coefficient which is nonzero. Choose a sequence, $z_n \to 0$, such that $a_k z_n^k < 0$. For example, let $-a_k z_n^k = (1/n)$. Then

$$|q(z_n)| = |1 + a_k z_n^k + r(z)| \leq 1 - 1/n + |r(z_n)| = 1 + a_k z_n^k + |r(z_n)| < 1$$

for all $n$ large enough because $|r(z_n)|$ is small compared with $|a_k z_n^k|$ since it involves higher order terms. This is a contradiction. ■
9. For any $x$ and $y$, exactly one of the following must hold. Either $x = y$, $x < y$, or $x > y$ (trichotomy).

**Proof:** First consider the transitive law. Suppose $x < y$ and $y < z$. Why is $x < z$? In other words, why is $z - x \in \mathbb{R}^+$? It is because $z - x = (z - y) + (y - x)$ and both $z - y, y - x \in \mathbb{R}^+$. Thus by trichotomy above, $z - x \in \mathbb{R}^+$ and so $z > x$.

Next consider addition to an inequality. If $x < y$ why is $x + z < y + z$? It is because

\[(y + z) - (x + z) = (y + z) + (-1)(x + z) = y + (-1)x + z + (-1)z = y - x \in \mathbb{R}^+.
\]

Next consider multiplication. If $x \leq 0$ and $y \leq 0$, why is $xy \geq 0$? First note there is nothing to show if either $x$ or $y$ equal $0$ so assume this is not the case. By $-x > 0$ and $-y > 0$. Therefore, by and what was proved about $-x = (-1)x$,

\[(-x)(-y) = (-1)^2 xy \in \mathbb{R}^+.
\]

Is $(-1)^2 = 1$? If so the claim is proved. But $-(-1) = (-1)^2$ and $-(-1) = 1$ because $-1 + 1 = 0$.

Next consider if $x > 0$ why is $x^{-1} > 0$? By either $x^{-1} = 0$ or $-x^{-1} \in \mathbb{R}^+$. It can’t happen that $x^{-1} = 0$ because then you would have to have $1 = 0x$ and as was shown earlier, $0x = 0$. Therefore, consider the possibility that $-x^{-1} \in \mathbb{R}^+$. This can’t work either because then you would have $(-1)x^{-1}x = (-1)(1) = -1$ and it would follow from that $-1 \in \mathbb{R}^+$. But this is impossible because if $x \in \mathbb{R}^+$, then $(-1)x = -x \in \mathbb{R}^+$ and contradicts which states that either $-x$ or $x$ is in $\mathbb{R}^+$ but not both.

Next consider if $x < 0$, why is $x^{-1} < 0$? As before, $x^{-1} \neq 0$. If $x^{-1} > 0$, then as before,

\[-x(x^{-1}) = -1 \in \mathbb{R}^+
\]

which was just shown not to occur.

Next consider if $x < y$ why is $xz < yz$ if $z > 0$? This follows because

\[yz - xz = z(y - x) \in \mathbb{R}^+
\]

since both $z$ and $y - x \in \mathbb{R}^+$.

Next consider if $x < y$ and $z < 0$, why is $xz > yz$? This follows because

\[zx - yz = z(x - y) \in \mathbb{R}^+
\]

by what was proved in.

The last two claims are obvious and left for you.

### 1.12 Polynomials

It will be very important to be able to work with polynomials in certain parts of linear algebra to be presented later.

**Definition 1.12.1** A polynomial is an expression of the form $a_n\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0$, $a_n \neq 0$ where the $a_i$ come from a field of scalars. Two polynomials are equal means that the coefficients match for each power of $\lambda$. The degree of a polynomial is the largest power of $\lambda$. Thus the degree of the above polynomial is $n$. Addition of polynomials is defined in the usual way as is multiplication of two polynomials. The leading term in the above polynomial is $a_n\lambda^n$. The coefficient of the leading term is called the leading coefficient. It is called a monic polynomial if $a_n = 1$. 

Note that the degree of the zero polynomial is not defined in the above.

**Lemma 1.12.2** Let \( f ( \lambda ) \) and \( g ( \lambda ) \neq 0 \) be polynomials. Then there exist polynomials, \( q ( \lambda ) \) and \( r ( \lambda ) \) such that
\[
f ( \lambda ) = q ( \lambda ) g ( \lambda ) + r ( \lambda )
\]
where the degree of \( r ( \lambda ) \) is less than the degree of \( g ( \lambda ) \) or \( r ( \lambda ) = 0 \). These polynomials \( q ( \lambda ) \) and \( r ( \lambda ) \) are unique.

**Proof:** Suppose that \( f ( \lambda ) - q ( \lambda ) g ( \lambda ) \) is never equal to 0 for any \( q ( \lambda ) \). If it is, then the conclusion follows. Now suppose
\[
r ( \lambda ) = f ( \lambda ) - q ( \lambda ) g ( \lambda )
\]
and the degree of \( r ( \lambda ) \) is \( m \geq n \) where \( n \) is the degree of \( g ( \lambda ) \). Say the leading term of \( r ( \lambda ) \) is \( b \lambda^m \) while the leading term of \( g ( \lambda ) \) is \( d \lambda^n \). Then letting \( a = b / d \), \( a \lambda^{m-n} g ( \lambda ) \) has the same leading term as \( r ( \lambda ) \). Thus the degree of \( r_1 ( \lambda ) = r ( \lambda ) - a \lambda^{m-n} g ( \lambda ) \) is no more than \( m - 1 \). Then
\[
r_1 ( \lambda ) = f ( \lambda ) - ( q ( \lambda ) g ( \lambda ) + a \lambda^{m-n} g ( \lambda ) ) = f ( \lambda ) - \left( q ( \lambda ) + \frac{q_1 ( \lambda )}{q_1 ( \lambda )} \right) g ( \lambda )
\]
Denote by \( S \) the set of polynomials \( f ( \lambda ) - g ( \lambda ) l ( \lambda ) \). Out of all these polynomials, there exists one which has smallest degree \( r ( \lambda ) \). Let this take place when \( l ( \lambda ) = q ( \lambda ) \). Then by the above argument, the degree of \( r ( \lambda ) \) is less than the degree of \( g ( \lambda ) \). Otherwise, there is one which has smaller degree. Thus \( f ( \lambda ) = g ( \lambda ) q ( \lambda ) + r ( \lambda ) \).

As to uniqueness, if you have \( r ( \lambda ), \tilde{r} ( \lambda ), q ( \lambda ), \hat{q} ( \lambda ) \) which work, then you would have
\[
( \hat{q} ( \lambda ) - q ( \lambda ) ) g ( \lambda ) = r ( \lambda ) - \tilde{r} ( \lambda )
\]
Now if the polynomial on the right is not zero, then neither is the one on the left. Hence this would involve two polynomials which are equal although their degrees are different. This is impossible. Hence \( r ( \lambda ) = \tilde{r} ( \lambda ) \) and so, matching coefficients implies that \( \hat{q} ( \lambda ) = q ( \lambda ) \). ■

**Definition 1.12.3** A polynomial \( f \) is said to divide a polynomial \( g \) if \( g ( \lambda ) = f ( \lambda ) r ( \lambda ) \) for some polynomial \( r ( \lambda ) \). Let \( \{ \phi_i ( \lambda ) \} \) be a finite set of polynomials. The greatest common divisor will be the **monic** polynomial \( q ( \lambda ) \) such that \( q ( \lambda ) \) divides each \( \phi_i ( \lambda ) \) and if \( p ( \lambda ) \) divides each \( \phi_i ( \lambda ) \), then \( p ( \lambda ) \) divides \( q ( \lambda ) \). The finite set of polynomials \( \{ \phi_i \} \) is said to be **relatively prime** if their greatest common divisor is 1. A polynomial \( f ( \lambda ) \) is irreducible if there is no polynomial with coefficients in \( \mathbb{F} \) which divides it except nonzero scalar multiples of \( f ( \lambda ) \) and constants. In other words, it is not possible to write \( f ( \lambda ) = a ( \lambda ) b ( \lambda ) \) where each of \( a ( \lambda ) \), \( b ( \lambda ) \) have degree less than the degree of \( f ( \lambda ) \).

**Proposition 1.12.4** The greatest common divisor is unique.

**Proof:** Suppose both \( q ( \lambda ) \) and \( q' ( \lambda ) \) work. Then \( q ( \lambda ) \) divides \( q' ( \lambda ) \) and the other way around and so
\[
q' ( \lambda ) = q ( \lambda ) l ( \lambda ), q ( \lambda ) = l' ( \lambda ) q' ( \lambda )
\]
Therefore, the two must have the same degree. Hence \( l' ( \lambda ) \), \( l ( \lambda ) \) are both constants. However, this constant must be 1 because both \( q ( \lambda ) \) and \( q' ( \lambda ) \) are monic. ■

**Theorem 1.12.5** Let \( \{ \phi_i ( \lambda ) \} \) be polynomials, not all of which are zero polynomials. Then there exists a greatest common divisor and it equals the monic polynomial \( \psi ( \lambda ) \) of smallest degree such that there exist polynomials \( r_i ( \lambda ) \) satisfying
\[
\psi ( \lambda ) = \sum_{i=1}^{p} r_i ( \lambda ) \phi_i ( \lambda )
\]
Proof: Let $S$ denote the set of monic polynomials which are of the form

$$\sum_{i=1}^{p} r_i (\lambda) \phi_i (\lambda)$$

where $r_i (\lambda)$ is a polynomial. Then $S \neq \emptyset$ because some $\phi_i (\lambda) \neq 0$. Then let the $r_i$ be chosen such that the degree of the expression $\sum_{i=1}^{p} r_i (\lambda) \phi_i (\lambda)$ is as small as possible. Letting $\psi (\lambda)$ equal this sum, it remains to verify it is the greatest common divisor. First, does it divide each $\phi_i (\lambda)$? Suppose it fails to divide $\phi_1 (\lambda)$. Then by Lemma 1.12.6

$$\phi_1 (\lambda) = \psi (\lambda) l (\lambda) + r (\lambda)$$

where degree of $r (\lambda)$ is less than that of $\psi (\lambda)$. Then dividing $r (\lambda)$ by the leading coefficient if necessary and denoting the result by $\psi_1 (\lambda)$, it follows the degree of $\psi_1 (\lambda)$ is less than the degree of $\psi (\lambda)$ and $\psi_1 (\lambda)$ equals for some $a \in \mathbb{F}$

$$\psi_1 (\lambda) = (\phi_1 (\lambda) - \psi (\lambda) l (\lambda)) a$$

This is one of the polynomials in $S$. Therefore, $\psi (\lambda)$ does not have the smallest degree after all because the degree of $\psi_1 (\lambda)$ is smaller. This is a contradiction. Therefore, $\psi (\lambda)$ divides $\phi_1 (\lambda)$.

Similarly it divides all the other $\phi_i (\lambda)$.

If $p (\lambda)$ divides all the $\phi_i (\lambda)$, then it divides $\psi (\lambda)$ because of the formula for $\psi (\lambda)$ which equals $\sum_{i=1}^{p} r_i (\lambda) \phi_i (\lambda)$. Thus $\psi (\lambda)$ satisfies the condition to be the greatest common divisor. This shows the greatest common divisor exists and equals the above description of it. 

Lemma 1.12.6 Suppose $\phi (\lambda)$ and $\psi (\lambda)$ are monic polynomials which are irreducible and not equal. Then they are relatively prime.

Proof: Suppose $\eta (\lambda)$ is a nonconstant polynomial. If $\eta (\lambda)$ divides $\phi (\lambda)$, then since $\phi (\lambda)$ is irreducible, $\eta (\lambda)$ equals $a \phi (\lambda)$ for some $a \in \mathbb{F}$. If $\eta (\lambda)$ divides $\psi (\lambda)$ then it must be of the form $b \psi (\lambda)$ for some $b \in \mathbb{F}$ and so it follows

$$\psi (\lambda) = \frac{a}{b} \phi (\lambda)$$

but both $\psi (\lambda)$ and $\phi (\lambda)$ are monic polynomials which implies $a = b$ and so $\psi (\lambda) = \phi (\lambda)$. This is assumed not to happen. It follows the only polynomials which divide both $\psi (\lambda)$ and $\phi (\lambda)$ are constants and so the two polynomials are relatively prime. Thus a polynomial which divides them both must be a constant, and if it is monic, then it must be 1. Thus 1 is the greatest common divisor.

Lemma 1.12.7 Let $\psi (\lambda)$ be an irreducible monic polynomial not equal to 1 which divides

$$\prod_{i=1}^{p} \phi_i (\lambda)^{k_i}, \; k_i \text{ a positive integer},$$

where each $\phi_i (\lambda)$ is an irreducible monic polynomial not equal to 1. Then $\psi (\lambda)$ equals some $\phi_i (\lambda)$.
Proof: Say \( \psi(\lambda) l(\lambda) = \prod_{i=1}^{p} \phi_{i}(\lambda)^{k_{i}} \). Suppose \( \psi(\lambda) \neq \phi_{i}(\lambda) \) for all \( i \). Then by Lemma 1.12.9, there exist polynomials \( m_{i}(\lambda), n_{i}(\lambda) \) such that

\[
1 = \psi(\lambda) m_{i}(\lambda) + \phi_{i}(\lambda) n_{i}(\lambda)
\]

\[
\phi_{i}(\lambda) n_{i}(\lambda) = 1 - \psi(\lambda) m_{i}(\lambda)
\]

Hence, \( (\phi_{i}(\lambda) n_{i}(\lambda))^{k_{i}} = (1 - \psi(\lambda) m_{i}(\lambda))^{k_{i}} \) and so letting \( n(\lambda) = \prod_{i} n_{i}(\lambda)^{k_{i}} \),

\[
n(\lambda) l(\lambda) \psi(\lambda) = \prod_{i=1}^{p} (n_{i}(\lambda) \phi_{i}(\lambda))^{k_{i}} = \prod_{i=1}^{p} (1 - \psi(\lambda) m_{i}(\lambda))^{k_{i}}
\]

\[
= 1 + g(\lambda) \psi(\lambda)
\]

for a suitable polynomial \( g(\lambda) \). You just separate out the term \( 1^{k_{i}} = 1 \) in that product and then all terms that are left have a \( \psi(\lambda) \) as a factor. Hence

\[
(n(\lambda) l(\lambda) - g(\lambda)) \psi(\lambda) = 1
\]

which is impossible because \( \psi(\lambda) \) is not equal to 1. \( \Box \)

Of course, since coefficients are in a field, you can drop the stipulation that the polynomials are monic and replace the conclusion with: \( \psi(\lambda) \) is a multiple of some \( \phi_{i}(\lambda) \).

Here now is a simple lemma about canceling monic polynomials.

Lemma 1.12.8 Suppose \( p(\lambda) \) is a monic polynomial and \( q(\lambda) \) is a polynomial such that

\[
p(\lambda) q(\lambda) = 0.
\]

Then \( q(\lambda) = 0 \). Also if

\[
p(\lambda) q_{1}(\lambda) = p(\lambda) q_{2}(\lambda)
\]

then \( q_{1}(\lambda) = q_{2}(\lambda) \).

Proof: Let

\[
p(\lambda) = \sum_{j=1}^{k} p_{j}\lambda^{j}, \quad q(\lambda) = \sum_{i=1}^{n} q_{i}\lambda^{i}, \quad p_{k} = 1.
\]

Then the product equals

\[
\sum_{j=1}^{k} \sum_{i=1}^{n} p_{j}q_{i}\lambda^{i+j}.
\]

If not all \( q_{i} = 0 \), let \( q_{m} \) be the last coefficient which is nonzero. Then the above is of the form

\[
\sum_{j=1}^{k} \sum_{i=1}^{m} p_{j}q_{i}\lambda^{i+j} = 0
\]

Consider the \( \lambda^{m+k} \) term. There is only one and it is \( p_{k}q_{m}\lambda^{m+k} \). Since \( p_{k} = 1, q_{m} = 0 \) after all. The second part follows from

\[
p(\lambda)(q_{1}(\lambda) - q_{2}(\lambda)) = 0. \quad \Box
\]

The following is the analog of the fundamental theorem of arithmetic for polynomials.

Theorem 1.12.9 Let \( f(\lambda) \) be a nonconstant polynomial with coefficients in \( \mathbb{F} \). Then there is some \( a \in \mathbb{F} \) such that \( f(\lambda) = a \prod_{i=1}^{n} \phi_{i}(\lambda) \) where \( \phi_{i}(\lambda) \) is an irreducible nonconstant monic polynomial and repeats are allowed. Furthermore, this factorization is unique in the sense that any two of these factorizations have the same nonconstant factors in the product, possibly in different order and the same constant \( a \).
Proof: That such a factorization exists is obvious. If \( f(\lambda) \) is irreducible, you are done. Factor out the leading coefficient. If not, then \( f(\lambda) = a\phi_1(\lambda)\phi_2(\lambda) \) where these are monic polynomials. Continue doing this with the \( \phi_i \) and eventually arrive at a factorization of the desired form.

It remains to argue the factorization is unique except for order of the factors. Suppose

\[
a \prod_{i=1}^{n} \phi_i(\lambda) = b \prod_{i=1}^{m} \psi_i(\lambda)
\]

where the \( \phi_i(\lambda) \) and the \( \psi_i(\lambda) \) are all irreducible monic nonconstant polynomials and \( a, b \in \mathbb{F} \). If \( n > m \), then by Lemma 1.12.7, each \( \psi_i(\lambda) \) equals one of the \( \phi_j(\lambda) \). By the above cancellation lemma, Lemma 1.12.8, you can cancel all these \( \psi_i(\lambda) \) with appropriate \( \phi_j(\lambda) \) and obtain a contradiction because the resulting polynomials on either side would have different degrees. Similarly, it cannot happen that \( n < m \). It follows \( n = m \) and the two products consist of the same polynomials. Then it follows \( a = b \).

The following corollary will be well used. This corollary seems rather believable but does require a proof.

**Corollary 1.12.10** Let \( q(\lambda) = \prod_{i=1}^{p} \phi_i(\lambda)^{k_i} \) where the \( k_i \) are positive integers and the \( \phi_i(\lambda) \) are irreducible distinct monic polynomials. Suppose also that \( p(\lambda) \) is a monic polynomial which divides \( q(\lambda) \). Then

\[
p(\lambda) = \prod_{i=1}^{p} \phi_i(\lambda)^{r_i}
\]

where \( r_i \) is a nonnegative integer no larger than \( k_i \).

Proof: Using Theorem 1.12.9, let \( p(\lambda) = b\prod_{i=1}^{s} \psi_i(\lambda)^{r_i} \) where the \( \psi_i(\lambda) \) are each irreducible and monic and \( b \in \mathbb{F} \). Since \( p(\lambda) \) is monic, \( b = 1 \). Then there exists a polynomial \( g(\lambda) \) such that

\[
p(\lambda) g(\lambda) = g(\lambda) \prod_{i=1}^{s} \psi_i(\lambda)^{r_i} = \prod_{i=1}^{p} \phi_i(\lambda)^{k_i}
\]

Hence \( g(\lambda) \) must be monic. Therefore,

\[
p(\lambda) g(\lambda) = \prod_{i=1}^{s} \psi_i(\lambda)^{r_i} \prod_{j=1}^{l} \eta_j(\lambda) = \prod_{i=1}^{p} \phi_i(\lambda)^{k_i}
\]

for \( \eta_j \) monic and irreducible. By uniqueness, each \( \psi_i \) equals one of the \( \phi_j(\lambda) \) and the same holding true of the \( \eta_j(\lambda) \). Therefore, \( p(\lambda) \) is of the desired form because you can cancel the \( \eta_j(\lambda) \) from both sides.

**1.13 Examples Of Finite Fields**

The emphasis of the first part of this book will be on what can be done on the basis of algebra alone. Linear algebra only needs a field of scalars along with some axioms involving an Abelian group of vectors and there are infinitely many examples of fields, including some which are finite. Since it is good to have examples in mind, I will present the finite fields of residue classes modulo a prime number in this little section. Then, when linear algebra is developed in the first part of the book and reference is made to a field of scalars, you should think that it is possible that the field might be this field of residue classes.
1.13.1 Division Of Numbers

First of all, recall the Archimedean property of the real numbers which says that if \( x \) is any real number, and if \( a > 0 \) then there exists a positive integer \( n \) such that \( na > x \). Geometrically, it is essentially the following: For any \( a > 0 \), the succession of disjoint intervals \([0, a), [a, 2a), [2a, 3a), \ldots\) includes all nonnegative real numbers. Here is a picture of some of these intervals.

\[
\begin{array}{cccc}
0a & 1a & 2a & 3a & 4a \\
\hline
\end{array}
\]

Then the version of the Euclidean algorithm presented here says that, for an arbitrary nonnegative real number \( b \), it is in exactly one interval

\[ [pa, (p + 1)a) \]

where \( p \) is some nonnegative integer. This seems obvious from the picture, but here is a proof.

**Theorem 1.13.1** Suppose \( 0 < a \) and let \( b \geq 0 \). Then there exists a unique integer \( p \) and real number \( r \) such that \( 0 \leq r < a \) and \( b = pa + r \).

**Proof:** Let \( S = \{n \in \mathbb{N} : an > b\} \). By the Archimedean property this set is nonempty. Let \( p + 1 \) be the smallest element of \( S \). Then \( pa \leq b \) because \( p + 1 \) is the smallest in \( S \). Therefore,

\[ r \equiv b - pa \geq 0. \]

If \( r \geq a \) then \( b - pa \geq a \) and so \( b \geq (p + 1)a \) contradicting \( p + 1 \in S \). Therefore, \( r < a \) as desired.

To verify uniqueness of \( p \) and \( r \), suppose \( p_1 \) and \( r_1 \), \( i = 1, 2 \), both work and \( r_2 > r_1 \). Then a little algebra shows

\[ p_1 - p_2 = \frac{r_2 - r_1}{a} \in (0, 1). \]

Thus \( p_1 - p_2 \) is an integer between 0 and 1, and there are no such integers. The case that \( r_1 > r_2 \) cannot occur either by similar reasoning. Thus \( r_1 = r_2 \) and it follows that \( p_1 = p_2 \). ■

**Corollary 1.13.2** The same conclusion is reached if \( b < 0 \).

**Proof:** In this case, \( S = \{n \in \mathbb{N} : an > -b\} \). Let \( p + 1 \) be the smallest element of \( S \). Then \( pa \leq -b < (p + 1)a \) and so \((-p) a \geq b > -(p + 1) a \). Let \( r \equiv b + (p + 1) \). Then \( b = - (p + 1)a + r \) and \( a > r \geq 0 \). As to uniqueness, say \( r_1 \) works and \( r_1 > r_2 \). Then you would have

\[ \begin{align*}
 b &= p_1 a + r_1 \\
 b &= p_2 a + r_2
\end{align*} \]

and

\[ p_2 - p_1 = \frac{r_1 - r_2}{a} \in (0, 1) \]

which is impossible because \( p_2 - p_1 \) is an integer. Hence \( r_1 = r_2 \) and so also \( p_1 = p_2 \). ■

**Corollary 1.13.3** Suppose \( a, b \neq 0 \), then there exists \( r \) such that \(|r| < |a|\) and for some \( p \) an integer,

\[ b = ap + r \]

**Proof:** This is done in the above except for the case where \( a < 0 \). So suppose this is the case. Then \( b = p(-a) + r \) where \( r \) is positive and \( 0 \leq r < -a = |a| \). Thus \( b = (-p)a + r \) such that \( 0 \leq |r| < |a| \). ■

This theorem is called the Euclidean algorithm when \( a \) and \( b \) are integers and this is the case of most interest here. Note that if \( a, b \) are integers, then so is \( r \). Note that

\[ 7 = 2 	imes 3 + 1, \quad 7 = 3 	imes 3 - 2, \quad |1| < 3, |{-2}| < 3 \]

so in this last corollary, the \( p \) and \( r \) are not unique.

The following definition describes what is meant by a prime number and also what is meant by the word “divides”.

Definition 1.13.4 The number, \(a\) divides the number, \(b\) if in Theorem 1.13.3, \(r = 0\). That is there is zero remainder. The notation for this is \(a \mid b\), read \(a\) divides \(b\) and \(a\) is called a factor of \(b\). A prime number is one which has the property that the only numbers which divide it are itself and 1. The greatest common divisor of two positive integers, \(m, n\) is that number, \(p\) which has the property that \(p\) divides both \(m\) and \(n\) and also if \(q\) divides both \(m\) and \(n\), then \(q\) divides \(p\). Two integers are relatively prime if their greatest common divisor is one. The greatest common divisor of \(m\) and \(n\) is denoted as \((m, n)\).

There is a phenomenal and amazing theorem which relates the greatest common divisor to the smallest number in a certain set. Suppose \(m, n\) are two positive integers. Then if \(x, y\) are integers, so is \(xm + yn\). Consider all integers which are of this form. Some are positive such as \(1m + 1n\) and some are not. The set \(S\) in the following theorem consists of exactly those integers of this form which are positive. Then the greatest common divisor of \(m\) and \(n\) will be the smallest number in \(S\). This is what the following theorem says.

Theorem 1.13.5 Let \(m, n\) be two positive integers and define

\[ S = \{ xm + yn \in \mathbb{N} : x, y \in \mathbb{Z} \}. \]

Then the smallest number in \(S\) is the greatest common divisor, denoted by \((m, n)\).

Proof: First note that both \(m\) and \(n\) are in \(S\) so it is a nonempty set of positive integers. By well ordering, there is a smallest element of \(S\), called \(p = x_0m + y_0n\). Either \(p\) divides \(m\) or it does not. If \(p\) does not divide \(m\), then by Theorem 1.13.3,

\[ m = pq + r \]

where \(0 < r < p\). Thus \(m = (x_0m + y_0n)q + r\) and so, solving for \(r\),

\[ r = m(1 - x_0) + (-y_0q) n \in S. \]

However, this is a contradiction because \(p\) was the smallest element of \(S\). Thus \(p \mid m\). Similarly \(p \mid n\).

Now suppose \(q\) divides both \(m\) and \(n\). Then \(m = qx\) and \(n = qy\) for integers, \(x\) and \(y\). Therefore,

\[ p = mx_0 + ny_0 = x_0qx + y_0qy = q(x_0x + y_0y) \]

showing \(q \mid p\). Therefore, \(p = (m, n)\). ■

This amazing theorem will now be used to prove a fundamental property of prime numbers which leads to the fundamental theorem of arithmetic, the major theorem which says every integer can be factored as a product of primes.

Theorem 1.13.6 If \(p\) is a prime and \(p \mid ab\) then either \(p \mid a\) or \(p \mid b\).

Proof: Suppose \(p\) does not divide \(a\). Then since \(p\) is prime, the only factors of \(p\) are 1 and \(p\) so follows \((p, a) = 1\) and therefore, there exists integers, \(x\) and \(y\) such that

\[ 1 = ax + yp. \]

Multiplying this equation by \(b\) yields

\[ b = abx + ybp. \]

Since \(p \mid ab\), \(ab = pz\) for some integer \(z\). Therefore,

\[ b = abx + ybp = pzx + ybp = p(xz + yb) \]

and this shows \(p\) divides \(b\). ■

Theorem 1.13.7 (Fundamental theorem of arithmetic) Let \(a \in \mathbb{N} \setminus \{1\}\). Then \(a = \prod_{i=1}^{n} p_i\) where \(p_i\) are all prime numbers. Furthermore, this prime factorization is unique except for the order of the factors.
1.13. EXAMPLES OF FINITE FIELDS

**Proof:** If \( a \) equals a prime number, the prime factorization clearly exists. In particular the prime factorization exists for the prime number 2. Assume this theorem is true for all \( a \leq n - 1 \). If \( n \) is a prime, then it has a prime factorization. On the other hand, if \( n \) is not a prime, then there exist two integers \( k \) and \( m \) such that \( n = km \) where each of \( k \) and \( m \) are less than \( n \). Therefore, each of these is no larger than \( n - 1 \) and consequently, each has a prime factorization. Thus so does \( n \). It remains to argue the prime factorization is unique except for order of the factors.

Suppose

\[
\prod_{i=1}^{n} p_i = \prod_{j=1}^{m} q_j
\]

where the \( p_i \) and \( q_j \) are all prime, there is no way to reorder the \( q_j \) such that \( m = n \) and \( p_i = q_j \) for all \( i \), and \( n + m \) is the smallest positive integer such that this happens. Then by Theorem 1.13.6, \( p_1|q_j \) for some \( j \). Since these are prime numbers this requires \( p_1 = q_j \). Reordering if necessary it can be assumed that \( q_j = q_1 \). Then dividing both sides by \( p_1 = q_1 \),

\[
\prod_{i=1}^{n-1} p_i + 1 = \prod_{j=1}^{m-1} q_j + 1.
\]

Since \( n + m \) was as small as possible for the theorem to fail, it follows that \( n - 1 = m - 1 \) and the prime numbers, \( q_2, \ldots, q_m \) can be reordered in such a way that \( p_k = q_k \) for all \( k = 2, \ldots, n \). Hence \( p_i = q_i \) for all \( i \) because it was already argued that \( p_1 = q_1 \), and this results in a contradiction. \( \blacksquare \)

1.13.2 The Field \( \mathbb{Z}_p \)

With this preparation, here is the construction of the finite fields \( \mathbb{Z}_p \) for \( p \) a prime.

**Definition 1.13.8** Let \( \mathbb{Z}^+ \) denote the set of nonnegative integers. Thus \( \mathbb{Z}^+ = \{0, 1, 2, 3, \ldots \} \). Also let \( p \) be a prime number. We will say that two integers, \( a, b \) are equivalent and write \( a \sim b \) if \( a - b \) is divisible by \( p \). Thus they are equivalent if \( a - b = px \) for some integer \( x \).

**Proposition 1.13.9** The relation \( \sim \) is an equivalence relation. Denoting by \( \bar{n} \) the equivalence class determined by \( n \in \mathbb{N} \), the following are well defined operations.

\[
\bar{n} + \bar{m} \equiv \bar{n + m}
\]

\[
\bar{n}\bar{m} \equiv \overline{nm}
\]

which makes the set \( \mathbb{Z}_p \) consisting of \( \{0, 1, \ldots, p - 1\} \) into a field.

**Proof:** First note that for \( n \in \mathbb{Z}^+ \) there always exists \( r \in \{0, 1, \ldots, p - 1\} \) such that \( \bar{n} = \bar{r} \). This is clearly true because if \( n \in \mathbb{Z}^+ \), then \( n = mp + r \) for \( r < p \), this by the Euclidean algorithm. Thus \( r = \bar{n} \).

Now suppose that \( n_1 = \bar{n} \) and \( m_1 = \bar{m} \). Is it true that \( \bar{n} + \bar{m} = \bar{n + m} \)? Is it true that \( \bar{n + m} = \bar{n + m} \)? Is \( \bar{n + m} \) a multiple of \( p \)? Of course since \( n_1 - n \) and \( m_1 - m \) are both multiples of \( p \). Similarly, is \( \overline{nm} \) a multiple of \( p \)? Of course this is so because

\[
nm - n_1 m_1 = nm - n_1 m + n_1 m - n_1 m_1
\]

\[
= m(n - n_1) + n_1(m - m_1)
\]

which is a multiple of \( p \). Thus the operations are well defined. It follows that all of the field axioms hold except possibly the existence of a multiplicative inverse and an additive inverse. First consider the question of an additive inverse. A typical thing in \( \mathbb{Z}_p \) is of the form \( \bar{r} \) where \( 0 \leq r \leq p - 1 \). Then consider \( (\bar{p} - \bar{r}) \). By definition, \( \bar{r} + \bar{p} = \bar{0} \) and so the additive inverse exists.

Now consider the existence of a multiplicative inverse. This is where \( p \) is prime is used. Say \( \bar{n} \neq \bar{0} \). That is, \( n \) is not a multiple of \( p \), \( 0 \leq n < p \). Then since \( p \) is prime, \( n, p \) are relatively prime and so there are integers \( x, y \) such that

\[
xn + yp = 1
\]
Choose \( m \geq 0 \) such that \( pm + x > 0, pm + y > 0 \). Then
\[
1 + pmn + pmp = (pm + x)n + (pm + y)p
\]
It follows that
\[
1 + pmn + p^2m = 1
\]
and so \((pm + x)\) is the multiplicative inverse of \( n \).

Thus \( \mathbb{Z}_p \) is a finite field, known as the field of residue classes modulo \( p \).

Something else which is often considered is a commutative ring with unity.

**Definition 1.13.10** A commutative ring with unity is just a field except it lacks the property that nonzero elements have a multiplicative inverse. It has all other properties. Thus the axioms of a commutative ring with unity are as follows:

**Axiom 1.13.11** Here are the axioms for a commutative ring with unity.

1. \( x + y = y + x \), (commutative law for addition)
2. There exists \( 0 \) such that \( x + 0 = x \) for all \( x \), (additive identity).
3. For each \( x \in \mathbb{F} \), there exists \( -x \in \mathbb{F} \) such that \( x + (-x) = 0 \), (existence of additive inverse).
4. \((x + y) + z = x + (y + z)\), (associative law for addition).
5. \( xy = yx \), (commutative law for multiplication). You could write this as \( x \times y = y \times x \).
6. \((xy)z = x(yz)\), (associative law for multiplication).
7. There exists \( 1 \) such that \( 1x = x \) for all \( x \), (multiplicative identity).
8. \( x(y + z) = xy + xz \), (distributive law).

An example of such a thing is \( \mathbb{Z}_m \) where \( m \) is not prime, also the ordinary integers. However, the integers are also an integral domain.

**Definition 1.13.12** A commutative ring with unity is called an integral domain if, in addition to the above, whenever \( ab = 0 \), it follows that either \( a = 0 \) or \( b = 0 \).

### 1.14 Some Topics From Analysis

Recall from calculus that if \( A \) is a nonempty set, \( \sup_{a \in A} f(a) \) denotes the least upper bound of \( f(A) \) or if this set is not bounded above, it equals \( \infty \). Also \( \inf_{a \in A} f(a) \) denotes the greatest lower bound of \( f(A) \) if this set is bounded below and it equals \( -\infty \) if \( f(A) \) is not bounded below. Thus to say \( \sup_{a \in A} f(a) = \infty \) is just a way to say that \( A \) is not bounded above.

**Definition 1.14.1** Let \( f(a, b) \in [-\infty, \infty] \) for \( a \in A \) and \( b \in B \) where \( A, B \) are sets which means that \( f(a, b) \) is either a number, \( \infty \), or \( -\infty \). The symbol, \( +\infty \) is interpreted as a point out at the end of the number line which is larger than every real number. Of course there is no such number. That is why it is called \( \infty \). The symbol, \( -\infty \) is interpreted similarly. Then \( \sup_{a \in A} f(a, b) \) means \( \sup (S_b) \) where \( S_b \equiv \{ f(a, b) : a \in A \} \).

Unlike limits, you can take the sup in different orders.

**Lemma 1.14.2** Let \( f(a, b) \in [-\infty, \infty] \) for \( a \in A \) and \( b \in B \) where \( A, B \) are sets. Then
\[
\sup_{a \in A} \sup_{b \in B} f(a, b) = \sup_{b \in B} \sup_{a \in A} f(a, b).
\]
Let denote by \( \lim \) the lower bound of \( f \) in the usual way and equals the least upper bound of \( f \). Therefore, if \( a \) is an increasing sequence of points of \( A \), then \( \lim \) does not exist. This is because the terms of the sequence are a distance of 1 apart. Therefore there can’t exist a single number such that all the terms of the sequence are ultimately within 1 of it. The nice thing about \( \lim \sup \) and \( \lim \inf \) is that they always exist. First here is a simple lemma and definition.

**Theorem 1.14.3** Let \( a_{ij} \geq 0 \). Then

\[
\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij}.
\]

**Proof:** First note there is no trouble in defining these sums because the \( a_{ij} \) are all nonnegative. If a sum diverges, it only diverges to \( \infty \) and so \( \infty \) is the value of the sum. Next note that

\[
\sum_{j=r}^{\infty} \sum_{i=r}^{\infty} a_{ij} \geq \sup_{n} \sum_{j=r}^{n} \sum_{i=r}^{n} a_{ij}
\]

because for all \( j \),

\[
\sum_{i=r}^{\infty} a_{ij} \geq \sum_{i=r}^{n} a_{ij}.
\]

Therefore,

\[
\sum_{j=r}^{\infty} \sum_{i=r}^{\infty} a_{ij} \geq \sup_{n} \sum_{j=r}^{n} \sum_{i=r}^{n} a_{ij} = \sup_{n} \lim_{m \to \infty} \sum_{j=r}^{m} \sum_{i=r}^{n} a_{ij}
\]

\[
= \sup_{n} \lim_{m \to \infty} \sum_{i=r}^{n} \sum_{j=r}^{m} a_{ij} = \sup_{n} \lim_{m \to \infty} \sum_{j=r}^{n} \sum_{i=r}^{m} a_{ij}
\]

\[
= \sup_{n} \sum_{i=r}^{\infty} \sum_{j=r}^{\infty} a_{ij} = \lim_{n \to \infty} \sum_{i=r}^{\infty} \sum_{j=r}^{\infty} a_{ij} = \sum_{i=r}^{\infty} \sum_{j=r}^{\infty} a_{ij}
\]

Interchanging the \( i \) and \( j \) in the above argument proves the theorem. ■

### 1.14.1 \( \lim \sup \) and \( \lim \inf \)

Sometimes the limit of a sequence does not exist. For example, if \( a_n = (-1)^n \), then \( \lim_{n \to \infty} a_n \) does not exist. This is because the terms of the sequence are a distance of 1 apart. Therefore there can’t exist a single number such that all the terms of the sequence are ultimately within 1/4 of that number. The nice thing about \( \lim \sup \) and \( \lim \inf \) is that they always exist. First here is a simple lemma and definition.

**Definition 1.14.4** Denote by \([- \infty, \infty]\) the real line along with symbols \( \infty \) and \(- \infty \). It is understood that \( \infty \) is larger than every real number and \(- \infty \) is smaller than every real number. Then if \( \{ A_n \} \) is an increasing sequence of points of \([- \infty, \infty]\), \( \lim_{n \to \infty} A_n \) equals \( \infty \) if the only upper bound of the set \( \{ A_n \} \) is \( \infty \). If \( \{ A_n \} \) is bounded above by a real number, then \( \lim_{n \to \infty} A_n \) is defined in the usual way and equals the least upper bound of \( \{ A_n \} \). If \( \{ A_n \} \) is a decreasing sequence of points of \([- \infty, \infty]\), \( \lim_{n \to \infty} A_n \) equals \(- \infty \) if the only lower bound of the sequence \( \{ A_n \} \) is \(- \infty \). If \( \{ A_n \} \) is bounded below by a real number, then \( \lim_{n \to \infty} A_n \) is defined in the usual way and equals the greatest lower bound of \( \{ A_n \} \). More simply, if \( \{ A_n \} \) is increasing,

\[
\lim_{n \to \infty} A_n = \sup \{ A_n \}
\]

and if \( \{ A_n \} \) is decreasing then

\[
\lim_{n \to \infty} A_n = \inf \{ A_n \}.
\]
Lemma 1.14.5 Let \( \{a_n\} \) be a sequence of real numbers and let \( U_n = \sup \{a_k : k \geq n\} \). Then \( \{U_n\} \) is a decreasing sequence. Also if \( L_n = \inf \{a_k : k \geq n\} \), then \( \{L_n\} \) is an increasing sequence. Therefore, \( \lim_{n \to \infty} L_n \) and \( \lim_{n \to \infty} U_n \) both exist.

**Proof:** Let \( W_n \) be an upper bound for \( \{a_k : k \geq n\} \). Then since these sets are getting smaller, it follows that for \( m < n \), \( W_m \) is an upper bound for \( \{a_k : k \geq n\} \). In particular if \( W_m = U_m \), then \( U_m \) is an upper bound for \( \{a_k : k \geq n\} \) and so \( U_m \) is at least as large as \( U_n \), the least upper bound for \( \{a_k : k \geq n\} \). The claim that \( \{L_n\} \) is decreasing is similar.

From the lemma, the following definition makes sense.

**Definition 1.14.6** Let \( \{a_n\} \) be any sequence of points in \([-\infty, \infty] \)

\[
\limsup_{n \to \infty} a_n = \lim \sup_{n \to \infty} \{a_k : k \geq n\}
\]

\[
\liminf_{n \to \infty} a_n = \lim \inf_{n \to \infty} \{a_k : k \geq n\}.
\]

**Theorem 1.14.7** Suppose \( \{a_n\} \) is a sequence of real numbers and that \( \limsup_{n \to \infty} a_n \) and \( \liminf_{n \to \infty} a_n \) are both real numbers. Then \( \lim_{n \to \infty} a_n \) exists if and only if

\[
\liminf_{n \to \infty} a_n = \limsup_{n \to \infty} a_n
\]

and in this case,

\[
\lim_{n \to \infty} a_n = \liminf_{n \to \infty} a_n = \limsup_{n \to \infty} a_n.
\]

**Proof:** First note that

\[
\sup_{n \to \infty} \{a_k : k \geq n\} \geq \inf \{a_k : k \geq n\}
\]

and so,

\[
\limsup_{n \to \infty} a_n = \lim \sup_{n \to \infty} \{a_k : k \geq n\} \geq \lim \inf_{n \to \infty} \{a_k : k \geq n\} = \liminf_{n \to \infty} a_n.
\]

Suppose first that \( \lim_{n \to \infty} a_n \) exists and is a real number \( a \). Then from the definition of a limit, there exists \( N \) corresponding to \( \varepsilon/6 \) in the definition. Hence, if \( m, n \geq N \), then

\[
|a_n - a_m| \leq |a_n - a| + |a - a_m| < \frac{\varepsilon}{6} + \frac{\varepsilon}{6} = \frac{\varepsilon}{3}.
\]

From the definition of \( \sup \{a_k : k \geq N\} \), there exists \( n_1 \geq N \) such that

\[
\sup \{a_k : k \geq N\} \leq a_{n_1} + \varepsilon/3.
\]

Similarly, there exists \( n_2 \geq N \) such that

\[
\inf \{a_k : k \geq N\} \geq a_{n_2} - \varepsilon/3.
\]

It follows that

\[
\sup \{a_k : k \geq N\} - \inf \{a_k : k \geq N\} \leq |a_{n_1} - a_{n_2}| + \frac{2\varepsilon}{3} < \varepsilon.
\]
Since the sequence, \( \{ \sup \{ a_k : k \geq N \} \}_{N=1}^{\infty} \) is decreasing and \( \{ \inf \{ a_k : k \geq N \} \}_{N=1}^{\infty} \) is increasing, it follows that

\[
0 \leq \lim_{N \to \infty} \sup \{ a_k : k \geq N \} - \lim_{N \to \infty} \inf \{ a_k : k \geq N \} \leq \varepsilon
\]

Since \( \varepsilon \) is arbitrary, this shows

\[
\lim_{N \to \infty} \sup \{ a_k : k \geq N \} = \lim_{N \to \infty} \inf \{ a_k : k \geq N \}
\]  

(1.1)

Next suppose and both equal \( a \in \mathbb{R} \). Then

\[
\lim_{N \to \infty} (\sup \{ a_k : k \geq N \} - \inf \{ a_k : k \geq N \}) = 0
\]

Since \( \sup \{ a_k : k \geq N \} \geq \inf \{ a_k : k \geq N \} \) it follows that for every \( \varepsilon > 0 \), there exists \( N \) such that

\[
\sup \{ a_k : k \geq N \} - \inf \{ a_k : k \geq N \} < \varepsilon,
\]

and for every \( N \),

\[
\inf \{ a_k : k \geq N \} \leq a \leq \sup \{ a_k : k \geq N \}
\]

Thus if \( n \geq N \),

\[
|a - a_n| < \varepsilon
\]

which implies that \( \lim_{n \to \infty} a_n = a \). In case

\[
a = \infty = \lim_{N \to \infty} \sup \{ a_k : k \geq N \} = \lim_{N \to \infty} \inf \{ a_k : k \geq N \}
\]

then if \( r \in \mathbb{R} \) is given, there exists \( N \) such that \( \inf \{ a_k : k \geq N \} > r \) which is to say that \( \lim_{n \to \infty} a_n = \infty \). The case where \( a = -\infty \) is similar except you use \( \sup \{ a_k : k \geq N \} \).

The significance of \( \lim \sup \) and \( \lim \inf \), in addition to what was just discussed, is contained in the following theorem which follows quickly from the definition.

**Theorem 1.14.8** Suppose \( \{a_n\} \) is a sequence of points of \([ -\infty, \infty ]\). Let

\[
\lambda = \lim_{n \to \infty} \sup a_n.
\]

Then if \( b > \lambda \), it follows there exists \( N \) such that whenever \( n \geq N \),

\[
a_n \leq b.
\]

If \( c < \lambda \), then \( a_n > c \) for infinitely many values of \( n \). Let

\[
\gamma = \lim_{n \to \infty} \inf a_n.
\]

Then if \( d < \gamma \), it follows there exists \( N \) such that whenever \( n \geq N \),

\[
a_n \geq d.
\]

If \( e > \gamma \), it follows \( a_n < e \) for infinitely many values of \( n \).

The proof of this theorem is left as an exercise for you. It follows directly from the definition and it is the sort of thing you must do yourself. Here is one other simple proposition.

**Proposition 1.14.9** Let \( \lim_{n \to \infty} a_n = a > 0 \). Then

\[
\lim_{n \to \infty} \sup a_n b_n = a \lim_{n \to \infty} b_n.
\]
**Proof:** This follows from the definition. Let \( \lambda_n = \sup \{a_k b_k : k \geq n\} \). For all \( n \) large enough, \( a_n > a - \varepsilon \) where \( \varepsilon \) is small enough that \( a - \varepsilon > 0 \). Therefore,

\[
\lambda_n \geq \sup \{b_k : k \geq n\} (a - \varepsilon)
\]

for all \( n \) large enough. Then

\[
\limsup_{n \to \infty} a_n b_n = \lim_{n \to \infty} \lambda_n = \limsup_{n \to \infty} a_n b_n \\
\geq \lim_{n \to \infty} \left( \sup \{b_k : k \geq n\} (a - \varepsilon) \right) \\
= (a - \varepsilon) \limsup_{n \to \infty} b_n
\]

Similar reasoning shows

\[
\limsup_{n \to \infty} a_n b_n \leq (a + \varepsilon) \limsup_{n \to \infty} b_n
\]

Now since \( \varepsilon > 0 \) is arbitrary, the conclusion follows.■

### 1.15 Exercises

1. Prove by induction that \( \sum_{k=1}^{n} k^3 = \frac{1}{4} n^4 + \frac{1}{2} n^3 + \frac{1}{4} n^2 \).  
2. Prove by induction that whenever \( n \geq 2, \sum_{k=1}^{n} \frac{1}{\sqrt{k}} > \sqrt{n} \).  
3. Prove by induction that \( 1 + \sum_{i=1}^{n} i \cdot (i!) = (n + 1)! \).  
4. The binomial theorem states \( (x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^{k} \) where

\[
\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1} \text{ if } k \in [1, n], \quad \binom{n}{0} = 1 = \binom{n}{n}
\]

Prove the binomial theorem by induction. Next show that

\[
\binom{n}{k} = \frac{n!}{(n-k)!k!}, \quad 0! = 1
\]

5. Let \( z = 5 + i9 \). Find \( z^{-1} \).  
6. Let \( z = 2 + i7 \) and let \( w = 3 - i8 \). Find \( zw, z + w, z^2 \), and \( w/z \).  
7. Give the complete solution to \( x^4 + 16 = 0 \).  
8. Graph the complex cube roots of 8 in the complex plane. Do the same for the four fourth roots of 16.  
9. If \( z \) is a complex number, show there exists \( \omega \) a complex number with \( |\omega| = 1 \) and \( \omega z = |z| \).  
10. De Moivre’s theorem says \( [r (\cos t + i \sin t)]^n = r^n (\cos nt + i \sin nt) \) for \( n \) a positive integer. Does this formula continue to hold for all integers \( n \), even negative integers? Explain.  
11. You already know formulas for \( \cos (x + y) \) and \( \sin (x + y) \) and these were used to prove De Moivre’s theorem. Now using De Moivre’s theorem, derive a formula for \( \sin (5x) \) and one for \( \cos (5x) \).  
12. If \( z \) and \( w \) are two complex numbers and the polar form of \( z \) involves the angle \( \theta \) while the polar form of \( w \) involves the angle \( \phi \), show that in the polar form for \( zw \) the angle involved is \( \theta + \phi \). Also, show that in the polar form of a complex number \( z, r = |z| \).
13. Factor $x^3 + 8$ as a product of linear factors.

14. Write $x^3 + 27$ in the form $(x + 3) (x^2 + ax + b)$ where $x^2 + ax + b$ cannot be factored any more using only real numbers.

15. Completely factor $x^4 + 16$ as a product of linear factors.

16. Factor $x^4 + 16$ as the product of two quadratic polynomials each of which cannot be factored further without using complex numbers.

17. If $z, w$ are complex numbers prove $zw = z w$ and then show by induction that $\prod_{j=1}^{n} z_j = \prod_{j=1}^{n} \overline{z_j}$. Also verify that $\sum_{k=1}^{n} z_k = \sum_{k=1}^{n} \overline{z_k}$. In words this says the conjugate of a product equals the product of the conjugates and the conjugate of a sum equals the sum of the conjugates.

18. Suppose $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ where all the $a_k$ are real numbers. Suppose also that $p(z) = 0$ for some $z \in \mathbb{C}$. Show it follows that $p(\overline{z}) = 0$ also.

19. Show that $1 + i, 2 + i$ are the only two zeros to $p(x) = x^2 (3 + 2i) x + (1 + 3i)$ so the zeros do not necessarily come in conjugate pairs if the coefficients are not real.

20. I claim that $1 = -1$. Here is why.

$$-1 = i^2 = \sqrt{-1} \sqrt{-1} = \sqrt{(-1)^2} = \sqrt{1} = 1.$$ 

This is clearly a remarkable result but is there something wrong with it? If so, what is wrong?

21. De Moivre’s theorem is really a grand thing. I plan to use it now for rational exponents, not just integers.

$$1 = 1^{1/4} = (\cos 2\pi + i \sin 2\pi)^{1/4} = \cos (\pi/2) + i \sin (\pi/2) = i.$$ 

Therefore, squaring both sides it follows $1 = -1$ as in the previous problem. What does this tell you about De Moivre’s theorem? Is there a profound difference between raising numbers to integer powers and raising numbers to non integer powers?

22. Review Problem 10 at this point. Now here is another question: If $n$ is an integer, is it always true that $(\cos \theta - i \sin \theta)^n = \cos (n\theta) - i \sin (n\theta)$? Explain.

23. Suppose you have any polynomial in $\cos \theta$ and $\sin \theta$. By this I mean an expression of the form $\sum_{\alpha=0}^{m} \sum_{\beta=0}^{n} a_{\alpha \beta} \cos^\alpha \theta \sin^\beta \theta$ where $a_{\alpha \beta} \in \mathbb{C}$. Can this always be written in the form $\sum_{\gamma=-(n+m)}^{m+n} b_{\gamma} \cos \gamma \theta + \sum_{\tau=-(n+m)}^{n+m} c_{\tau} \sin \tau \theta$? Explain.

24. Suppose $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ is a polynomial and it has $n$ zeros, $z_1, z_2, \cdots, z_n$ listed according to multiplicity. ($z$ is a root of multiplicity $m$ if the polynomial $f(x) = (x - z)^m$ divides $p(x)$ but $(x - z) f(x)$ does not.) Show that

$$p(x) = a_n (x - z_1)(x - z_2) \cdots (x - z_n).$$

25. Give the solutions to the following quadratic equations having real coefficients.

(a) $x^2 - 2x + 2 = 0$

(b) $3x^2 + x + 3 = 0$
CHAPTER 1. SOME PREREQUISITE TOPICS

26. Give the solutions to the following quadratic equations having complex coefficients. Note how
the solutions do not come in conjugate pairs as they do when the equation has real coefficients.

(a) \( x^2 + 2x + 1 + i = 0 \)
(b) \( 4x^2 + 4ix - 5 = 0 \)
(c) \( 4x^2 + (4 + 4i)x + 1 + 2i = 0 \)
(d) \( x^2 - 4ix - 5 = 0 \)
(e) \( 3x^2 + (1 - i)x + 3i = 0 \)

27. Prove the fundamental theorem of algebra for quadratic polynomials having coefficients in \( \mathbb{C} \).
That is, show that an equation of the form \( ax^2 + bx + c = 0 \) where \( a, b, c \) are complex numbers,
\( a \neq 0 \) has a complex solution. \textbf{Hint:} Consider the fact, noted earlier that the expressions
given from the quadratic formula do in fact serve as solutions.

28. Does there exist a subset of \( \mathbb{C} \), \( \mathbb{C}^+ \) which satisfies \( 1.11.1 - 1.11.3 \)? \textbf{Hint:} You might review
the theorem about order. Show \(-1\) cannot be in \( \mathbb{C}^+ \). Now ask questions about \(-i\) and \(i\). In
mathematics, you can sometimes show certain things do not exist. It is very seldom you can
do this outside of mathematics. For example, does the Loch Ness monster exist? Can you
prove it does not?

29. Prove the Euclidean algorithm: If \( m, n \) are positive integers, then there exist integers \( q, r \geq 0 \)
such that \( r < m \) and

\[ n = qm + r \]

\textbf{Hint:} You might try considering

\[ S \equiv \{ n - km : k \in \mathbb{N} \text{ and } n - km < 0 \} \]

and picking the smallest integer in \( S \) or something like this. It was done in the chapter, but
go through it yourself.

30. \( \star \)Recall that two polynomials are equal means that the coefficients of corresponding powers
of \( \lambda \) are equal. Thus a polynomial equals 0 if and only if all coefficients equal 0. In calculus
we usually think of a polynomial as 0 if it sends every value of \( x \) to 0. Suppose you have the
following polynomial

\[ \overline{1}x^2 + \overline{1}x \]

where it is understood to be a polynomial in \( \mathbb{Z}_2 \). Thus it is not the zero polynomial. Show,
however, that this equals zero for all \( x \in \mathbb{Z}_2 \) so we would be tempted to say it is zero if we use
the conventions of calculus.

31. Prove Wilson’s theorem. This theorem states that if \( p \) is a prime, then \((p - 1)! + 1\) is divisible
by \( p \). Wilson’s theorem was first proved by Lagrange in the 1770’s. \textbf{Hint:} Check directly for
\( p = 2, 3 \). Show that \( \overline{p - 1} = -\overline{1} \) and that if \( a \in \{2, \ldots, p - 2\} \), then \( \overline{(a)}^{-1} \neq \overline{a} \). Thus a residue
class \( \overline{a} \) and its multiplicative inverse for \( a \in \{2, \ldots, p - 2\} \) occur in pairs. Show that this
implies that the residue class of \((p - 1)!\) must be \(-1\). From this, draw the conclusion.

32. \( \star \)Show that in the arithmetic of \( \mathbb{Z}_p \), \( (\overline{a} + \overline{b})^p = (\overline{a})^p + (\overline{b})^p \), a well known formula among
students.
33. Consider \((\pi) \in \mathbb{Z}_p\) for \(p\) a prime, and suppose \((\pi) \neq \overline{1,0}\). Fermat’s little theorem says that \((\pi)^{p-1} = \overline{1}\). In other words \((a)^{p-1} - 1\) is divisible by \(p\). Prove this. **Hint:** Show that there must exist \(r \geq 1, r \leq p - 1\) such that \((\pi)^r = \overline{1}\). To do so, consider \(\overline{1}, (\pi), (\pi)^2, \ldots\). Then these all have values in \(\{\overline{1}, \overline{2}, \ldots, \overline{p-1}\}\), and so there must be a repeat in \(\{\overline{1}, (\pi), (\pi)^2, \ldots, (\pi)^{p-1}\}\), say \(p - 1 \geq l > k\) and \((\pi)^l = (\pi)^k\). Then tell why \((\pi)^l-k - \overline{1} = 0\). Let \(r\) be the first positive integer such that \((\pi)^r = \overline{1}\). Let \(G = \{\overline{1}, (\pi), \ldots, (\pi)^{r-1}\}\). Show that every residue class in \(G\) has its multiplicative inverse in \(G\). In fact, \((\pi)^k (\pi)^{-k} = \overline{1}\). Also verify that the entries in \(G\) must be distinct. Now consider the sets \(\mathbb{Z}_p \times \mathbb{Z}_p \equiv \{b (\pi)^k : \overline{0}, \ldots, \overline{r-1}\}\) where \(b \in \{\overline{1}, \overline{2}, \ldots, \overline{p-1}\}\). Show that two of these sets are either the same or disjoint and that they all consist of \(r\) elements. Explain why it follows that \(p - 1 = \ell r\) for some positive integer \(\ell\) equal to the number of these distinct sets. Then explain why \((\pi)^{p-1} = (\pi)^R = \overline{1}\).

34. Let \(p(x)\) and \(q(x)\) be polynomials. Then by the division algorithm, there exist polynomials \(l(x), r(x)\) equal to 0 or having degree smaller than \(p(x)\) such that

\[q(x) = p(x) l(x) + r(x)\]

If \(k(x)\) is the greatest common divisor of \(p(x)\) and \(q(x)\), explain why \(k(x)\) must divide \(r(x)\). Then argue that \(k(x)\) is also the greatest common divisor of \(p(x)\) and \(r(x)\). Now repeat the process for the polynomials \(p(x)\) and \(r(x)\). This time, the remainder term will have degree smaller than \(r(x)\). Keep doing this and eventually the remainder must be 0. Describe an algorithm based on this which will determine the greatest common divisor of two polynomials.

35. Consider \(\mathbb{Z}_m\) where \(m\) is not a prime. Show that although this will not be a field, it is a commutative ring with unity.

36. This and the next few problems are to illustrate the utility of the \(\limsup\). A sequence of numbers \(\{x_n\}\) in \(\mathbb{C}\) is called a Cauchy sequence if for every \(\varepsilon > 0\) there exists \(m\) such that if \(k, l \geq m\), then \(|x_k - x_l| < \varepsilon\). The complex numbers are said to be complete because any Cauchy sequence converges. This is one form of the completeness axiom. Using this axiom, show that \(\sum_{k=0}^{\infty} r^k \equiv \lim_{n \to \infty} \sum_{k=0}^{n} r^k = \frac{1}{1-r}\) whenever \(r \in \mathbb{C}\) and \(|r| < 1\). **Hint:** You need to do a computation with the sum and show that the partial sums form a Cauchy sequence.

37. Show that if \(\sum_{j=1}^{\infty} |c_j|\) converges, meaning that \(\lim_{n \to \infty} \sum_{j=1}^{n} |c_j|\) exists, then \(\sum_{j=1}^{\infty} c_j\) also converges, meaning \(\lim_{n \to \infty} \sum_{j=1}^{n} c_j\) exists, this for \(c_j \in \mathbb{C}\). Recall from calculus, this says that absolute convergence implies convergence.

38. Show that if \(\sum_{j=1}^{\infty} c_j\) converges, meaning \(\lim_{n \to \infty} \sum_{j=1}^{n} c_j\) exists, then it must be the case that \(\lim_{n \to \infty} c_n = 0\).

39. Show that if \(\limsup_{k \to \infty} |a_k|^{1/k} < 1\), then \(\sum_{k=1}^{\infty} |a_k|\) converges, while if \(\limsup_{n \to \infty} |a_n|^{1/n} > 1\), then the series diverges spectacularly because \(\lim_{n \to \infty} |c_n|\) fails to equal 0 and in fact has a subsequence which converges to \(\infty\). Also show that if \(\limsup_{n \to \infty} |a_n|^{1/n} = 1\), the test fails because there are examples where the series can converge and examples where the series diverges. This is an improved version of the root test from calculus. It is improved because \(\limsup\) always exists. **Hint:** For the last part, consider \(\sum_{n} \frac{1}{n^2}\) and \(\sum_{n} \frac{1}{n}\). Review calculus to see why the first diverges and the second converges.

40. Consider a power series \(\sum_{n=0}^{\infty} a_n x^n\). Derive a condition for the radius of convergence using \(\limsup_{n \to \infty} |a_n|^{1/n}\). Recall that the radius of convergence \(R\) is such that if \(|x| < R\), then the series converges and if \(|x| > R\), the series diverges and if \(|x| = R\) it is not known whether the series converges. In this problem, assume only that \(x \in \mathbb{C}\).

41. Show that if \(a_n\) is a sequence of real numbers, then \(\liminf_{n \to \infty} (-a_n) = -\limsup_{n \to \infty} a_n\).
Part I

Linear Algebra For Its Own Sake
Chapter 2

Systems Of Linear Equations

This part of the book is about linear algebra itself, as a part of algebra. Some geometric and analytic concepts do creep in, but it is primarily about algebra. It involves general fields and has very little to do with limits and completeness although some geometry is included, but not much.

2.1 Elementary Operations

In this chapter, the main interest is in fields of scalars consisting of \( \mathbb{R} \) or \( \mathbb{C} \), but everything is applied to arbitrary fields. Consider the following example.

**Example 2.1.1** Find \( x \) and \( y \) such that

\[
x + y = 7 \quad \text{and} \quad 2x - y = 8.
\]

The set of ordered pairs, \((x, y)\) which solve both equations is called the **solution set**.

You can verify that \((5, 2)\) is a solution to the above system. The interesting question is this: If you were not given this information to verify, how could you determine the solution? You can do this by using the following basic operations on the equations, none of which change the set of solutions of the system of equations.

**Definition 2.1.2** Elementary operations are those operations consisting of the following.

1. Interchange the order in which the equations are listed.
2. Multiply any equation by a nonzero number.
3. Replace any equation with itself added to a multiple of another equation.

**Example 2.1.3** To illustrate the third of these operations on this particular system, consider the following.

\[
\begin{align*}
x + y &= 7 \\
2x - y &= 8
\end{align*}
\]

The system has the same solution set as the system

\[
\begin{align*}
x + y &= 7 \\
-3y &= -6
\end{align*}
\]

To obtain the second system, take the second equation of the first system and add \(-2\) times the first equation to obtain

\[-3y = -6.\]
Now, this clearly shows that \( y = 2 \) and so it follows from the other equation that \( x + 2 = 7 \) and so \( x = 5 \).

Of course a linear system may involve many equations and many variables. The solution set is still the collection of solutions to the equations. In every case, the above operations of Definition 2.1 are do not change the set of solutions to the system of linear equations.

**Theorem 2.1.4** Suppose you have two equations, involving the variables,

\[
(x_1, \cdots, x_n)
\]

\[
E_1 = f_1, E_2 = f_2
\]  

(2.2)

where \( E_1 \) and \( E_2 \) are expressions involving the variables and \( f_1 \) and \( f_2 \) are constants. (In the above example there are only two variables, \( x \) and \( y \) and \( E_1 = x + y \) while \( E_2 = 2x - y \).) Then the system \( E_1 = f_1, E_2 = f_2 \) has the same solution set as

\[
E_1 = f_1, E_2 + aE_1 = f_2 + af_1.
\]  

(2.3)

Also the system \( E_1 = f_1, E_2 = f_2 \) has the same solutions as the system, \( E_2 = f_2, E_1 = f_1 \). The system \( E_1 = f_1, E_2 = f_2 \) has the same solution as the system \( E_1 = f_1, aE_2 = af_2 \) provided \( a \neq 0 \).

**Proof:** If \((x_1, \cdots, x_n)\) solves \( E_1 = f_1, E_2 = f_2 \) then it solves the first equation in \( E_1 = f_1, E_2 + aE_1 = f_2 + af_1 \). Also, it satisfies \( aE_1 = af_1 \) and so, since it also solves \( E_2 = f_2 \) it must solve \( E_2 + aE_1 = f_2 + af_1 \). Therefore, if \((x_1, \cdots, x_n)\) solves \( E_1 = f_1, E_2 = f_2 \) it must also solve \( E_2 + aE_1 = f_2 + af_1 \). On the other hand, if it solves the system \( E_1 = f_1 \) and \( E_2 + aE_1 = f_2 + af_1 \), then \( aE_1 = af_1 \) and so you can subtract these equal quantities from both sides of \( E_2 + aE_1 = f_2 + af_1 \) to obtain \( E_2 = f_2 \) showing that it satisfies \( E_1 = f_1, E_2 = f_2 \).

The second assertion of the theorem which says that the system \( E_1 = f_1, E_2 = f_2 \) has the same solution as the system, \( E_2 = f_2, E_1 = f_1 \) is seen to be true because it involves nothing more than listing the two equations in a different order. They are the same equations.

The third assertion of the theorem which says \( E_1 = f_1, E_2 = f_2 \) is a solution as the system \( E_1 = f_1, aE_2 = af_2 \) if \( a \neq 0 \) is verified as follows: If \((x_1, \cdots, x_n)\) is a solution of \( E_1 = f_1, E_2 = f_2 \), then it is a solution to \( E_1 = f_1, aE_2 = af_2 \) because the second system only involves multiplying the equation, \( E_2 = f_2 \) by \( a \). If \((x_1, \cdots, x_n)\) is a solution of \( E_1 = f_1, aE_2 = af_2 \), then upon multiplying \( aE_2 = af_2 \) by the number \( 1/a \), you find that \( E_2 = f_2 \). 

Stated simply, the above theorem shows that the elementary operations do not change the solution set of a system of equations.

Here is an example in which there are three equations and three variables. You want to find values for \( x, y, z \) such that each of the given equations are satisfied when these values are plugged in to the equations.

**Example 2.1.5** Find the solutions to the system,

\[
x + 3y + 6z = 25
\]

\[
2x + 7y + 14z = 58
\]

\[
2y + 5z = 19
\]  

(2.4)

To solve this system replace the second equation by \((-2)\) times the first equation added to the second. This yields the system

\[
x + 3y + 6z = 25
\]

\[
y + 2z = 8
\]  

(2.5)

\[
2y + 5z = 19
\]

Now take \((-2)\) times the second and add to the third. More precisely, replace the third equation with \((-2)\) times the second added to the third. This yields the system

\[
x + 3y + 6z = 25
\]

\[
y + 2z = 8
\]

\[
z = 3
\]  

(2.6)
At this point, you can tell what the solution is. This system has the same solution as the original system and in the above, \( z = 3 \). Then using this in the second equation, it follows \( y + 6 = 8 \) and so \( y = 2 \). Now using this in the top equation yields \( x + 6 + 18 = 25 \) and so \( x = 1 \). This process is called back substitution.

Alternatively, in Exercise 2.6 you could have continued as follows. Add \((-2)\) times the bottom equation to the middle and then add \((-6)\) times the bottom to the top. This yields

\[
\begin{align*}
x + 3y &= 7 \\
y &= 2 \\
z &= 3
\end{align*}
\]

Now add \((-3)\) times the second to the top. This yields

\[
\begin{align*}
x &= 1 \\
y &= 2 \\
z &= 3
\end{align*}
\]

a system which has the same solution set as the original system. This avoided back substitution and led to the same solution set.

### 2.2 Gauss Elimination

A less cumbersome way to represent a linear system is to write it as an augmented matrix. For example the linear system, can be written as

\[
\begin{pmatrix}
1 & 3 & 6 & | & 25 \\
2 & 7 & 14 & | & 58 \\
0 & 2 & 5 & | & 19
\end{pmatrix}.
\]

It has exactly the same information as the original system but here it is understood there is an

\[
\begin{pmatrix}
1 \\
2 \\
0
\end{pmatrix}, \text{ a } \begin{pmatrix}
3 \\
7 \\
2
\end{pmatrix} \text{ and a } \begin{pmatrix}
6 \\
14 \\
5
\end{pmatrix}. \text{ The rows correspond to the}
\]

\[
\text{equations in the system. Thus the top row in the augmented matrix corresponds to the equation,}
\]

\[
x + 3y + 6z = 25.
\]

Now when you replace an equation with a multiple of another equation added to itself, you are just taking a row of this augmented matrix and replacing it with a multiple of another row added to it. Thus the first step in solving Exercise 2.4 would be to take \((-2)\) times the first row of the augmented matrix above and add it to the second row,

\[
\begin{pmatrix}
1 & 3 & 6 & | & 25 \\
0 & 1 & 2 & | & 8 \\
0 & 2 & 5 & | & 19
\end{pmatrix}.
\]

Note how this corresponds to Exercise 2.4. Next take \((-2)\) times the second row and add to the third,

\[
\begin{pmatrix}
1 & 3 & 6 & | & 25 \\
0 & 1 & 2 & | & 8 \\
0 & 0 & 1 & | & 3
\end{pmatrix}.
\]
CHAPTER 2. SYSTEMS OF LINEAR EQUATIONS

This augmented matrix corresponds to the system

\[
\begin{align*}
  x + 3y + 6z &= 25 \\
  y + 2z &= 8 \\
  z &= 3
\end{align*}
\]

which is the same as \ref{2.6}. By back substitution you obtain the solution \(x = 1, y = 6, \) and \(z = 3\).

In general a linear system is of the form

\[
\begin{align*}
  a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\
  \vdots \\
  a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m
\end{align*}
\]  

(2.7)

where the \(x_i\) are variables and the \(a_{ij}\) and \(b_i\) are constants. This system can be represented by the augmented matrix

\[
\begin{pmatrix}
  a_{11} & \cdots & a_{1n} & | & b_1 \\
  \vdots & & \vdots & | & \vdots \\
  a_{m1} & \cdots & a_{mn} & | & b_m
\end{pmatrix}
\]

(2.8)

Changes to the system of equations in \ref{2.7} as a result of an elementary operations translate into changes of the augmented matrix resulting from a row operation. Note that Theorem \ref{2.1.4} implies that the row operations deliver an augmented matrix for a system of equations which has the same solution set as the original system.

**Definition 2.2.1** The row operations consist of the following

1. Switch two rows.
2. Multiply a row by a nonzero number.
3. Replace a row by a multiple of another row added to it.

**Gauss elimination** is a systematic procedure to simplify an augmented matrix to a reduced form. In the following definition, the term “leading entry” refers to the first nonzero entry of a row when scanning the row from left to right.

**Definition 2.2.2** An augmented matrix is in **echelon form** if

1. All nonzero rows are above any rows of zeros.
2. Each leading entry of a row is in a column to the right of the leading entries of any rows above it.

How do you know when to stop doing row operations? You might stop when you have obtained an echelon form as described above, but you certainly should stop doing row operations if you have gotten a matrix in row reduced echelon form described next.

**Definition 2.2.3** An augmented matrix is in **row reduced echelon form** if

1. All nonzero rows are above any rows of zeros.
2. Each leading entry of a row is in a column to the right of the leading entries of any rows above it.
3. All entries in a column above and below a leading entry are zero.
4. Each leading entry is a 1, the only nonzero entry in its column.
Example 2.2.4 Here are some matrices which are in row reduced echelon form.

$$\begin{pmatrix}
1 & 0 & 0 & 5 & 8 & 0 \\
0 & 1 & 2 & 7 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}, \quad \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{pmatrix}.$$

Example 2.2.5 Here are matrices in echelon form which are not in row reduced echelon form but which are in echelon form.

$$\begin{pmatrix}
1 & 0 & 6 & 5 & 8 & 2 \\
0 & 2 & 2 & 7 & 3 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}, \quad \begin{pmatrix}
1 & 3 & 5 & 4 & 0 & 2 \\
0 & 2 & 0 & 7 & 0 & 0 \\
0 & 0 & 3 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0
\end{pmatrix}.$$

Example 2.2.6 Here are some matrices which are not in echelon form.

$$\begin{pmatrix}
0 & 0 & 0 & 0 \\
1 & 2 & 3 & 3 \\
0 & 1 & 0 & 2 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{pmatrix}, \quad \begin{pmatrix}
1 & 2 & 3 \\
2 & 4 & -6 \\
4 & 0 & 7 \\
0 & 0 & 1
\end{pmatrix}, \quad \begin{pmatrix}
0 & 2 & 3 & 3 \\
1 & 5 & 0 & 2 \\
7 & 5 & 0 & 1 \\
0 & 0 & 1 & 0
\end{pmatrix}.$$

Definition 2.2.7 A **pivot position** in a matrix is the location of a leading entry in an echelon form resulting from the application of row operations to the matrix. A **pivot column** is a column that contains a pivot position.

For example consider the following.

Example 2.2.8 Suppose

$$A = \begin{pmatrix}
1 & 2 & 3 & 4 \\
3 & 2 & 1 & 6 \\
4 & 4 & 4 & 10
\end{pmatrix}.$$  

Where are the pivot positions and pivot columns?

Replace the second row by $-3$ times the first added to the second. This yields

$$\begin{pmatrix}
1 & 2 & 3 & 4 \\
0 & -4 & -8 & -6 \\
4 & 4 & 4 & 10
\end{pmatrix}.$$  

This is not in reduced echelon form so replace the bottom row by $-4$ times the top row added to the bottom. This yields

$$\begin{pmatrix}
1 & 2 & 3 & 4 \\
0 & -4 & -8 & -6 \\
0 & -4 & -8 & -6
\end{pmatrix}.$$  

This is still not in reduced echelon form. Replace the bottom row by $-1$ times the middle row added to the bottom. This yields

$$\begin{pmatrix}
1 & 2 & 3 & 4 \\
0 & -4 & -8 & -6 \\
0 & 0 & 0 & 0
\end{pmatrix}.$$
which is in echelon form, although not in reduced echelon form. Therefore, the pivot positions in the original matrix are the locations corresponding to the first row and first column and the second row and second columns as shown in the following:

\[
\begin{bmatrix}
1 & 2 & 3 & 4 \\
3 & 2 & 1 & 6 \\
4 & 4 & 4 & 10
\end{bmatrix}
\]

Thus the pivot columns in the matrix are the first two columns.

The following is the algorithm for obtaining a matrix which is in row reduced echelon form.

**Algorithm 2.2.9**

This algorithm tells how to start with a matrix and do row operations on it in such a way as to end up with a matrix in row reduced echelon form.

1. Find the first nonzero column from the left. This is the first pivot column. The position at the top of the first pivot column is the first pivot position. Switch rows if necessary to place a nonzero number in the first pivot position.

2. Use row operations to zero out the entries below the first pivot position.

3. Ignore the row containing the most recent pivot position identified and the rows above it. Repeat steps 1 and 2 to the remaining sub-matrix, the rectangular array of numbers obtained from the original matrix by deleting the rows you just ignored. Repeat the process until there are no more rows to modify. The matrix will then be in echelon form.

4. Moving from right to left, use the nonzero elements in the pivot positions to zero out the elements in the pivot columns which are above the pivots.

5. Divide each nonzero row by the value of the leading entry. The result will be a matrix in row reduced echelon form.

This row reduction procedure applies to both augmented matrices and non augmented matrices. There is nothing special about the augmented column with respect to the row reduction procedure. From now on, I will not bother to include the line between the right column and what is before it.

**Example 2.2.10** Here is a matrix.

\[
\begin{bmatrix}
0 & 0 & 2 & 3 & 2 \\
0 & 1 & 1 & 4 & 3 \\
0 & 0 & 1 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 2 & 1
\end{bmatrix}
\]

Do row reductions till you obtain a matrix in echelon form. Then complete the process by producing one in row reduced echelon form.

The pivot column is the second. Hence the pivot position is the one in the first row and second column. Switch the first two rows to obtain a nonzero entry in this pivot position.
Step two is not necessary because all the entries below the first pivot position in the resulting matrix are zero. Now ignore the top row and the columns to the left of this first pivot position. Thus you apply the same operations to the smaller matrix

\[
\begin{pmatrix}
2 & 3 & 2 \\
1 & 2 & 2 \\
0 & 0 & 0 \\
0 & 2 & 1
\end{pmatrix}.
\]

The next pivot column is the third corresponding to the first in this smaller matrix and the second pivot position is therefore, the one which is in the second row and third column. In this case it is not necessary to switch any rows to place a nonzero entry in this position because there is already a nonzero entry there. Multiply the third row of the original matrix by \(-2\) and then add the second row to it. This yields

\[
\begin{pmatrix}
0 & 1 & 1 & 4 & 3 \\
0 & 0 & 2 & 3 & 2 \\
0 & 0 & 0 & -1 & -2 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 2 & 1
\end{pmatrix}.
\]

The next matrix the steps in the algorithm are applied to is

\[
\begin{pmatrix}
-1 & -2 \\
0 & 0 \\
2 & 1
\end{pmatrix}.
\]

The first pivot column is the first column in this case and no switching of rows is necessary because there is a nonzero entry in the first pivot position. Therefore, the algorithm yields for the next step

\[
\begin{pmatrix}
0 & 1 & 1 & 4 & 3 \\
0 & 0 & 2 & 3 & 2 \\
0 & 0 & 0 & -1 & -2 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -3
\end{pmatrix}.
\]

Now the algorithm will be applied to the matrix

\[
\begin{pmatrix}
0 \\
-3
\end{pmatrix}.
\]

There is only one column and it is nonzero so this single column is the pivot column. Therefore, the algorithm yields the following matrix for the echelon form.

\[
\begin{pmatrix}
0 & 1 & 1 & 4 & 3 \\
0 & 0 & 2 & 3 & 2 \\
0 & 0 & 0 & -1 & -2 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}.
\]

To complete placing the matrix in reduced echelon form, multiply the third row by 3 and add \(-2\)
times the fourth row to it. This yields
\[
\begin{pmatrix}
0 & 1 & 1 & 4 & 3 \\
0 & 0 & 2 & 3 & 2 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
Next multiply the second row by 3 and take 2 times the fourth row and add to it. Then add the fourth row to the first.
\[
\begin{pmatrix}
0 & 1 & 1 & 4 & 0 \\
0 & 0 & 6 & 9 & 0 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
Next work on the fourth column in the same way.
\[
\begin{pmatrix}
0 & 3 & 3 & 0 & 0 \\
0 & 0 & 6 & 0 & 0 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
Take \(-1/2\) times the second row and add to the first.
\[
\begin{pmatrix}
0 & 3 & 0 & 0 & 0 \\
0 & 0 & 6 & 0 & 0 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
Finally, divide by the value of the leading entries in the nonzero rows.
\[
\begin{pmatrix}
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
The above algorithm is the way a computer would obtain a reduced echelon form for a given matrix. It is not necessary for you to pretend you are a computer but if you like to do so, the algorithm described above will work. The main idea is to do row operations in such a way as to end up with a matrix in echelon form or row reduced echelon form because when this has been done, the resulting augmented matrix will allow you to describe the solutions to the linear system of equations in a meaningful way. When you do row operations until you obtain row reduced echelon form, the process is called the Gauss Jordan method. Otherwise, it is called Gauss elimination.

Example 2.2.11 Give the complete solution to the system of equations, \(5x + 10y - 7z = -2, 2x + 4y - 3z = -1,\) and \(3x + 6y + 5z = 9.\)
The augmented matrix for this system is

\[
\begin{pmatrix}
2 & 4 & -3 & -1 \\
5 & 10 & -7 & -2 \\
3 & 6 & 5 & 9
\end{pmatrix}
\]

Multiply the second row by 2, the first row by 5, and then take \((-1)\) times the first row and add to the second. Then multiply the first row by \(1/5\). This yields

\[
\begin{pmatrix}
2 & 4 & -3 & -1 \\
0 & 0 & 1 & 1 \\
3 & 6 & 5 & 9
\end{pmatrix}
\]

Now, combining some row operations, take \((-3)\) times the first row and add this to 2 times the last row and replace the last row with this. This yields.

\[
\begin{pmatrix}
2 & 4 & -3 & -1 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 21
\end{pmatrix}
\]

One more row operation, taking \((-1)\) times the second row and adding to the bottom yields.

\[
\begin{pmatrix}
2 & 4 & -3 & -1 \\
0 & 0 & 1 & 1 \\
0 & 0 & 0 & 20
\end{pmatrix}
\]

This is impossible because the last row indicates the need for a solution to the equation

\[
0x + 0y + 0z = 20
\]

and there is no such thing because \(0 \neq 20\). This shows there is no solution to the three given equations. When this happens, the system is called inconsistent. In this case it is very easy to describe the solution set. The system has no solution.

Here is another example based on the use of row operations.

**Example 2.2.12** Give the complete solution to the system of equations, \(3x - y - 5z = 9\), \(y - 10z = 0\), and \(-2x + y = -6\).

The augmented matrix of this system is

\[
\begin{pmatrix}
3 & -1 & -5 & 9 \\
0 & 1 & -10 & 0 \\
-2 & 1 & 0 & -6
\end{pmatrix}
\]

Replace the last row with 2 times the top row added to 3 times the bottom row combining two row operations. This gives

\[
\begin{pmatrix}
3 & -1 & -5 & 9 \\
0 & 1 & -10 & 0 \\
0 & 1 & -10 & 0
\end{pmatrix}
\]

The entry, 3 in this sequence of row operations is called the pivot. It is used to create zeros in the other places of the column. Next take \(-1\) times the middle row and add to the bottom. Here the 1 in the second row is the pivot.

\[
\begin{pmatrix}
3 & -1 & -5 & 9 \\
0 & 1 & -10 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]
Take the middle row and add to the top and then divide the top row which results by 3.

\[
\begin{pmatrix}
1 & 0 & -5 & 3 \\
0 & 1 & -10 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}.
\]

This is in reduced echelon form. The equations corresponding to this reduced echelon form are

\[
y = 10z \quad \text{and} \quad x = 3 + 5z.
\]

Apparently \( z \) can equal any number. Let's call this number \( t \). Therefore, the solution set of this system is \( x = 3 + 5t, y = 10t, \) and \( z = t \) where \( t \) is completely arbitrary. The system has an infinite set of solutions which are given in the above simple way. This is what it is all about, finding the solutions to the system.

There is some terminology connected to this which is useful. Recall how each column corresponds to a variable in the original system of equations. The variables corresponding to a pivot column are called basic variables. The other variables are called free variables. In Example 2.2.12 there was one free variable, \( z \), and two basic variables, \( x \) and \( y \). In describing the solution to the system of equations, the free variables are assigned a parameter. In Example 2.2.12 this parameter was \( t \). Sometimes there are many free variables and in these cases, you need to use many parameters. Here is another example.

**Example 2.2.13** Find the solution to the system

\[
\begin{align*}
x + 2y - z + w &= 3 \\
x + y - z + w &= 1 \\
x + 3y - z + w &= 5
\end{align*}
\]

The augmented matrix is

\[
\begin{pmatrix}
1 & 2 & -1 & 1 & 3 \\
1 & 1 & -1 & 1 & 1 \\
1 & 3 & -1 & 1 & 5
\end{pmatrix}.
\]

Take \(-1\) times the first row and add to the second. Then take \(-1\) times the first row and add to the third. This yields

\[
\begin{pmatrix}
1 & 2 & -1 & 1 & 3 \\
0 & -1 & 0 & 0 & -2 \\
0 & 1 & 0 & 0 & 2
\end{pmatrix}
\]

Now add the second row to the bottom row

\[
\begin{pmatrix}
1 & 2 & -1 & 1 & 3 \\
0 & -1 & 0 & 0 & -2 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix} \quad \text{(2.9)}
\]

This matrix is in echelon form and you see the basic variables are \( x \) and \( y \) while the free variables are \( z \) and \( w \). Assign \( s \) to \( z \) and \( t \) to \( w \). Then the second row yields the equation, \( y = 2 \) while the top equation yields the equation, \( x + 2y - s + t = 3 \) and so since \( y = 2 \), this gives \( x + 4 - s + t = 3 \) showing that \( x = -1 + s - t, y = 2, z = s, \) and \( w = t \). It is customary to write this in the form

\[
\begin{pmatrix}
x \\
y \\
z \\
w
\end{pmatrix} = \begin{pmatrix}
-1 + s - t \\
2 \\
s \\
t
\end{pmatrix}.
\]

\[\text{(2.10)}\]

\[\text{1In this context } t \text{ is called a parameter.}\]
This is another example of a system which has an infinite solution set but this time the solution set depends on two parameters, not one. Most people find it less confusing in the case of an infinite solution set to first place the augmented matrix in row reduced echelon form rather than just echelon form before seeking to write down the description of the solution. In the above, this means we don’t stop with the echelon form \[ 2.10 \]. Instead we first place it in reduced echelon form as follows.

\[
\begin{pmatrix}
1 & 0 & -1 & 1 & -1 \\
0 & 1 & 0 & 0 & 2 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

Then the solution is \( y = 2 \) from the second row and \( x = -1 + z - w \) from the first. Thus letting \( z = s \) and \( w = t \), the solution is given in \[ 2.10 \).

The number of free variables is always equal to the number of different parameters used to describe the solution. If there are no free variables, then either there is no solution as in the case where row operations yield an echelon form like

\[
\begin{pmatrix}
1 & 2 & 3 \\
0 & 4 & -2 \\
0 & 0 & 1
\end{pmatrix}
\]

or there is a unique solution as in the case where row operations yield an echelon form like

\[
\begin{pmatrix}
1 & 2 & 2 & 3 \\
0 & 4 & 3 & -2 \\
0 & 0 & 4 & 1
\end{pmatrix}
\]

Also, sometimes there are free variables and no solution as in the following:

\[
\begin{pmatrix}
1 & 2 & 2 & 3 \\
0 & 4 & 3 & -2 \\
0 & 0 & 0 & 1
\end{pmatrix}
\]

There are a lot of cases to consider but it is not necessary to make a major production of this. Do row operations till you obtain a matrix in echelon form or reduced echelon form and determine whether there is a solution. If there is, see if there are free variables. In this case, there will be infinitely many solutions. Find them by assigning different parameters to the free variables and obtain the solution. If there are no free variables, then there will be a unique solution which is easily determined once the augmented matrix is in echelon or row reduced echelon form. In every case, the process yields a straightforward way to describe the solutions to the linear system. As indicated above, you are probably less likely to become confused if you place the augmented matrix in row reduced echelon form rather than just echelon form.

In summary,

**Definition 2.2.14** A system of linear equations is a list of equations,

\[
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\
\vdots \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m
\]

where \( a_{ij} \) are numbers, and \( b_j \) is a number. The above is a system of \( m \) equations in the \( n \) variables, \( x_1, x_2, \ldots, x_n \). Nothing is said about the relative size of \( m \) and \( n \). Written more simply in terms of summation notation, the above can be written in the form

\[
\sum_{j=1}^{n} a_{ij}x_j = f_i, \quad i = 1, 2, 3, \ldots, m
\]
It is desired to find \((x_1, \cdots, x_n)\) solving each of the equations listed.

As illustrated above, such a system of linear equations may have a unique solution, no solution, or infinitely many solutions and these are the only three cases which can occur for any linear system. Furthermore, you do exactly the same things to solve any linear system. You write the augmented matrix and do row operations until you get a simpler system in which it is possible to see the solution, usually obtaining a matrix in echelon or reduced echelon form. All is based on the observation that the row operations do not change the solution set. You can have more equations than variables, fewer equations than variables, etc. It doesn’t matter. You always set up the augmented matrix and go to work on it.

**Definition 2.2.15** A system of linear equations is called **consistent** if there exists a solution. It is called **inconsistent** if there is no solution.

These are reasonable words to describe the situations of having or not having a solution. If you think of each equation as a condition which must be satisfied by the variables, consistent would mean there is some choice of variables which can satisfy all the conditions. Inconsistent would mean there is no choice of the variables which can satisfy each of the conditions.

### 2.3 Exercises

1. Here is an augmented matrix in which \(\ast\) denotes an arbitrary number and \(\blacksquare\) denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?
   
   \[
   \begin{pmatrix}
   \blacksquare & \ast & \ast & \ast & \ast & | & \ast \\
   0 & \blacksquare & \ast & \ast & 0 & | & \ast \\
   0 & 0 & \blacksquare & \ast & \ast & | & \ast \\
   0 & 0 & 0 & 0 & \blacksquare & | & \ast
   \end{pmatrix}
   \]

2. Here is an augmented matrix in which \(\ast\) denotes an arbitrary number and \(\blacksquare\) denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?
   
   \[
   \begin{pmatrix}
   \blacksquare & \ast & \ast & | & \ast \\
   0 & \blacksquare & \ast & | & \ast \\
   0 & 0 & \blacksquare & | & \ast
   \end{pmatrix}
   \]

3. Here is an augmented matrix in which \(\ast\) denotes an arbitrary number and \(\blacksquare\) denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?
   
   \[
   \begin{pmatrix}
   \blacksquare & \ast & \ast & \ast & \ast & | & \ast \\
   0 & \blacksquare & 0 & \ast & 0 & | & \ast \\
   0 & 0 & 0 & \blacksquare & \ast & | & \ast \\
   0 & 0 & 0 & 0 & \blacksquare & | & \ast
   \end{pmatrix}
   \]

4. Here is an augmented matrix in which \(\ast\) denotes an arbitrary number and \(\blacksquare\) denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?
   
   \[
   \begin{pmatrix}
   \blacksquare & \ast & \ast & \ast & \ast & | & \ast \\
   0 & \blacksquare & \ast & \ast & 0 & | & \ast \\
   0 & 0 & 0 & \blacksquare & | & 0 \\
   0 & 0 & 0 & 0 & \ast & | & \blacksquare
   \end{pmatrix}
   \]
5. Suppose a system of equations has fewer equations than variables. Must such a system be consistent? If so, explain why and if not, give an example which is not consistent.

6. If a system of equations has more equations than variables, can it have a solution? If so, give an example and if not, tell why not.

7. Find $h$ such that
\[
\begin{pmatrix}
2 & h & | & 4 \\
3 & 6 & | & 7
\end{pmatrix}
\]
is the augmented matrix of an inconsistent matrix.

8. Find $h$ such that
\[
\begin{pmatrix}
1 & h & | & 3 \\
2 & 4 & | & 6
\end{pmatrix}
\]
is the augmented matrix of a consistent matrix.

9. Find $h$ such that
\[
\begin{pmatrix}
1 & 1 & | & 4 \\
3 & h & | & 12
\end{pmatrix}
\]
is the augmented matrix of a consistent matrix.

10. Choose $h$ and $k$ such that the augmented matrix shown has one solution. Then choose $h$ and $k$ such that the system has no solutions. Finally, choose $h$ and $k$ such that the system has infinitely many solutions.
\[
\begin{pmatrix}
1 & h & | & 2 \\
2 & 4 & | & k
\end{pmatrix}
\]

11. Choose $h$ and $k$ such that the augmented matrix shown has one solution. Then choose $h$ and $k$ such that the system has no solutions. Finally, choose $h$ and $k$ such that the system has infinitely many solutions.
\[
\begin{pmatrix}
1 & 2 & | & 2 \\
2 & h & | & k
\end{pmatrix}
\]

12. Determine if the system is consistent. If so, is the solution unique?
\[
x + 2y + z - w = 2 \\
x - y + z + w = 1 \\
2x + y - z = 1 \\
4x + 2y + z = 5
\]

13. Determine if the system is consistent. If so, is the solution unique?
\[
x + 2y + z - w = 2 \\
x - y + z + w = 0 \\
2x + y - z = 1 \\
4x + 2y + z = 3
\]

14. Find the general solution of the system whose augmented matrix is
\[
\begin{pmatrix}
1 & 2 & 0 & | & 2 \\
1 & 3 & 4 & | & 2 \\
1 & 0 & 2 & | & 1
\end{pmatrix}
\]
15. Find the general solution of the system whose augmented matrix is
\[
\begin{pmatrix}
1 & 2 & 0 & | & 2 \\
2 & 0 & 1 & | & 1 \\
3 & 2 & 1 & | & 3
\end{pmatrix}
\]

16. Find the general solution of the system whose augmented matrix is
\[
\begin{pmatrix}
1 & 1 & 0 & | & 1 \\
1 & 0 & 4 & | & 2
\end{pmatrix}
\]

17. Find the general solution of the system whose augmented matrix is
\[
\begin{pmatrix}
1 & 0 & 2 & 1 & 1 & | & 2 \\
0 & 1 & 0 & 1 & 2 & | & 1 \\
1 & 2 & 0 & 0 & 1 & | & 3 \\
1 & 0 & 1 & 0 & 2 & | & 2
\end{pmatrix}
\]

18. Find the general solution of the system whose augmented matrix is
\[
\begin{pmatrix}
1 & 0 & 2 & 1 & 1 & | & 2 \\
0 & 1 & 0 & 1 & 2 & | & 1 \\
0 & 2 & 0 & 0 & 1 & | & 3 \\
1 & -1 & 2 & 2 & 2 & | & 0
\end{pmatrix}
\]

19. Give the complete solution to the system of equations, \(7x + 14y + 15z = 22\), \(2x + 4y + 3z = 5\), and \(3x + 6y + 10z = 13\).

20. Give the complete solution to the system of equations, \(3x - y + 4z = 6\), \(y + 8z = 0\), and \(-2x + y = -4\).

21. Give the complete solution to the system of equations, \(9x - 2y + 4z = -17\), \(13x - 3y + 6z = -25\), and \(-2x - z = 3\).

22. Give the complete solution to the system of equations, \(65x + 84y + 16z = 546\), \(81x + 105y + 20z = 682\), and \(84x + 110y + 21z = 713\).

23. Give the complete solution to the system of equations, \(8x + 2y + 3z = -3\), \(8x + 3y + 3z = -1\), and \(4x + y + 3z = -9\).

24. Give the complete solution to the system of equations, \(-8x + 2y + 5z = 18\), \(-8x + 3y + 5z = 13\), and \(-4x + y + 5z = 19\).

25. Give the complete solution to the system of equations, \(3x - y - 2z = 3\), \(y - 4z = 0\), and \(-2x + y = -2\).

26. Give the complete solution to the system of equations, \(-9x + 15y = 66\), \(-11x + 18y = 79\), \(-x + y = 4\), and \(z = 3\).

27. Give the complete solution to the system of equations, \(-19x + 8y = -108\), \(-71x + 30y = -404\), \(-2x + y = -12\), \(4x + z = 14\).

28. Consider the system \(-5x + 2y - z = 0\) and \(-5x - 2y - z = 0\). Both equations equal zero and so \(-5x + 2y - z = -5x - 2y - z\) which is equivalent to \(y = 0\). Thus \(x\) and \(z\) can equal anything. But when \(x = 1\), \(z = -4\), and \(y = 0\) are plugged in to the equations, it doesn’t work. Why?
29. Four times the weight of Gaston is 150 pounds more than the weight of Ichabod. Four times the weight of Ichabod is 660 pounds less than seventeen times the weight of Gaston. Four times the weight of Gaston plus the weight of Siegfried equals 290 pounds. Brunhilde would balance all three of the others. Find the weights of the four sisters.

30. The steady state temperature, \( u \) in a plate solves Laplace’s equation, \( \Delta u = 0 \). One way to approximate the solution which is often used is to divide the plate into a square mesh and require the temperature at each node to equal the average of the temperature at the four adjacent nodes. This procedure is justified by the mean value property of harmonic functions. In the following picture, the numbers represent the observed temperature at the indicated nodes. Your task is to find the temperature at the interior nodes, indicated by \( x, y, z, \) and \( w \). One of the equations is \( z = \frac{1}{4} (10 + 0 + w + x) \).

31. Consider the following diagram of four circuits.

Those jagged places denote resistors and the numbers next to them give their resistance in ohms, written as \( \Omega \). The breaks in the lines having one short line and one long line denote a voltage source which causes the current to flow in the direction which goes from the longer of the two lines toward the shorter along the unbroken part of the circuit. The current in amps in the four circuits is denoted by \( I_1, I_2, I_3, I_4 \) and it is understood that the motion is in the counter clockwise direction. If \( I_k \) ends up being negative, then it just means the current flows in the clockwise direction. Then Kirchhoff’s law states that

The sum of the resistance times the amps in the counter clockwise direction around a loop equals the sum of the voltage sources in the same direction around the loop.

In the above diagram, the top left circuit should give the equation

\[
2I_2 - 2I_1 + 5I_2 - 5I_3 + 3I_2 = 5
\]

For the circuit on the lower left, you should have

\[
4I_1 + I_1 - I_4 + 2I_1 - 2I_2 = -10
\]

Write equations for each of the other two circuits and then give a solution to the resulting system of equations. You might use a computer algebra system to find the solution. It might be more convenient than doing it by hand.
32. Consider the following diagram of three circuits.

Those jagged places denote resistors and the numbers next to them give their resistance in ohms, written as Ω. The breaks in the lines having one short line and one long line denote a voltage source which causes the current to flow in the direction which goes from the longer of the two lines toward the shorter along the unbroken part of the circuit. The current in amps in the four circuits is denoted by $I_1, I_2, I_3$ and it is understood that the motion is in the counter clockwise direction. If $I_k$ ends up being negative, then it just means the current flows in the clockwise direction. Then Kirchhoff’s law states that the sum of the resistance times the amps in the counter clockwise direction around a loop equals the sum of the voltage sources in the same direction around the loop. Find $I_1, I_2, I_3$. 

Chapter 3

Vector Spaces

It is time to consider the idea of an abstract Vector space which is something which has two operations satisfying the following vector space axioms.

**Definition 3.0.1** A vector space is an Abelian group of “vectors” satisfying the axioms of an Abelian group,
\[ v + w = w + v, \]
the commutative law of addition,
\[ (v + w) + z = v + (w + z), \]
the associative law for addition,
\[ v + 0 = v, \]
the existence of an additive identity,
\[ v + (-v) = 0, \]
the existence of an additive inverse, along with a field of “scalars” \( \mathbb{F} \) which are allowed to multiply the vectors according to the following rules. (The Greek letters denote scalars.)
\[ \alpha (v + w) = \alpha v + \alpha v, \] (3.1)
\[ (\alpha + \beta) v = \alpha v + \beta v, \] (3.2)
\[ \alpha (\beta v) = \alpha \beta (v), \] (3.3)
\[ 1v = v. \] (3.4)

The field of scalars is usually \( \mathbb{R} \) or \( \mathbb{C} \) and the vector space will be called real or complex depending on whether the field is \( \mathbb{R} \) or \( \mathbb{C} \). However, other fields are also possible. For example, one could use the field of rational numbers or even the field of the integers mod \( p \) for \( p \) a prime. A vector space is also called a linear space. These axioms do not tell us anything about what is being considered. Nevertheless, one can prove some fundamental properties just based on these vector space axioms.

**Proposition 3.0.2** In any vector space, 0 is unique, \(-x\) is unique, \(0x = 0\), and \((-1)x = -x\).

**Proof:** Suppose \(0'\) is also an additive identity. Then for 0 the additive identity in the axioms,
\[ 0' = 0' + 0 = 0 \]
Next suppose \(x+y=0\). Then add \(-x\) to both sides.
\[ -x = -x + (x+y) = (-x+x) + y = 0 + y = y \]
Thus if \( y \) acts like the additive inverse, it is the additive inverse.

\[
0x = (0 + 0) x = 0x + 0x
\]

Now add \(-0x\) to both sides. This gives \( 0 = 0x \). Finally,

\[
(-1) x + x = (-1) x + 1x = (-1 + 1) x = 0x = 0
\]

By the uniqueness of the additive inverse shown earlier, \((-1) x = -x\). ■

If you are interested in considering other fields, you should have some examples other than \( \mathbb{C} \), \( \mathbb{R} \), \( \mathbb{Q} \). Some of these are discussed in the following exercises. If you are happy with only considering \( \mathbb{R} \) and \( \mathbb{C} \), skip these exercises. Here is an important example which gives the typical vector space.

**Example 3.0.3** Let \( \Omega \) be a nonempty set and define \( V \) to be the set of functions defined on \( \Omega \). Letting \( a, b, c \) be scalars and \( f, g, h \) functions, the vector operations are defined as

\[
(f + g)(x) = f(x) + g(x) \\
(af)(x) = a(f(x))
\]

Then this is an example of a vector space. Note that the set where the functions have their values can be any vector space.

To verify this, we check the axioms.

\[
(f + g)(x) = f(x) + g(x) = g(x) + f(x) = (g + f)(x)
\]

Since \( x \) is arbitrary, \( f + g = g + f \).

\[
((f + g) + h)(x) = (f + g)(x) + h(x) = (f(x) + g(x)) + h(x)
\]

\[
= f(x) + (g(x) + h(x)) = (f(x) + (g + h)(x)) = (f + (g + h))(x)
\]

and so \((f + g) + h = f + (g + h)\). Let \( 0 \) denote the function which is given by \( 0(x) = 0 \). Then this is an additive identity because

\[
(f + 0)(x) = f(x) + 0(x) = f(x)
\]

and so \( f + 0 = f \). Let \(-f\) be the function which satisfies \((-f)(x) = -f(x)\). Then

\[
(f + (-f))(x) = f(x) + (-f)(x) = f(x) + -f(x) = 0
\]

Hence \( f + (-f) = 0 \).

\[
((a + b)f)(x) = (a + b)f(x) = af(x) + bf(x) = (af + bf)(x)
\]

and so \((a + b)f = af + bf\).

\[
(a(f + g))(x) = a(f + g)(x) = a(f(x) + g(x))
\]

\[
= af(x) + ag(x) = (af + bg)(x)
\]

and so \(a(f + g) = af + bg\).

\[
((ab)f)(x) = (ab)f(x) = a(bf(x)) = (a(bf))(x)
\]

so \((abf) = a(bf)\). Finally \((1f)(x) = 1f(x) = f(x)\) so \(1f = f\).

As above, \( F \) will be a field.
3.1 Linear Combinations Of Vectors, Independence

The fundamental idea in linear algebra is the following notion of a linear combination.

**Definition 3.1.1** Let $x_1, \ldots, x_n$ be vectors in a vector space. A finite linear combination of these vectors is a vector which is of the form $\sum_{j=1}^n a_j x_j$ where the $a_j$ are scalars. In short, it is a sum of scalars times vectors. $\text{span}(x_1, \ldots, x_n)$ denotes the set of all linear combinations of the vectors $x_1, \ldots, x_n$. More generally, if $S$ is any set of vectors, $\text{span}(S)$ consists of all finite linear combinations of vectors from $S$.

**Definition 3.1.2** Let $(V, F)$ be a vector space and its field of scalars. Then $S \subseteq V$ is said to be linearly independent if whenever $\{v_1, \ldots, v_m\} \subseteq V$ with the $v_i$ distinct, then there is only one way to have a linear combination

$$\sum_{i=1}^n c_i v_i = 0$$

and this is to have each $c_i = 0$. More succinctly, if $\sum_{i=1}^n c_i v_i = 0$ then each $c_i = 0$. A set $S \subseteq V$ is linearly dependent if it is not linearly independent. That is, there is some subset of $S \{v_1, \ldots, v_n\}$ and scalars $c_i$ not all zero such that $\sum_{i=1}^n c_i v_i = 0$.

**Proposition 3.1.3** A set of vectors $\{u_1, \ldots, u_r\}$ is independent if and only if no vector is a linear combination of the others.

**Proof:** $\Rightarrow$ Suppose $\{u_1, \ldots, u_r\}$ is linearly independent. Could you have

$$u_i = \sum_{j \neq i} c_j u_j$$

No. This is not possible because if the above holds, then you would have

$$0 = (-1) u_i + \sum_{j \neq i} c_j u_j$$

in contradiction to the assumption that $\{u_1, \ldots, u_r\}$ is linearly independent.

$\Leftarrow$ Suppose now that no vector in the list is a linear combination of the others. Suppose $\sum_{i=1}^n c_i u_i = 0$. It is desired to show that whenever this happens, each $c_i = 0$. Could any of the $c_i$ be non zero? No. If $c_k \neq 0$, then you would have

$$\sum_{i=1}^n \frac{c_i}{c_k} u_i = 0$$
and so

\[ u_k = \sum_{i \neq k} -\frac{c_i}{c_k} u_i \]

showing that one can obtain \( u_k \) as a linear combination of the other vectors after all. It follows that all \( c_i = 0 \) and so \( \{u_1, \ldots, u_r\} \) is linearly independent. ■

**Example 3.1.4** Determine whether the real valued functions defined on \( \mathbb{R} \) given by the polynomials

\[ x^2 + 2x + 1, x^2 + 2x, x^2 + x + 1 \]

are independent.

Suppose

\[ a (x^2 + 2x + 1) + b (x^2 + 2x) + c (x^2 + x + 1) = 0 \]

then differentiate both sides to obtain

\[ a (2x + 2) + b (2x + 2) + c (2x + 1) = 0 \]

Now differentiate again.

\[ 2a + 2b + 2c = 0 \]

In the second equation, let \( x = -1 \). Then \(-c = 0\) so \( c = 0 \). Thus

\[
\begin{align*}
    a (x^2 + 2x + 1) + b (x^2 + 2x) &= 0 \\
    a + b &= 0
\end{align*}
\]

Now let \( x = 0 \) in the top equation to find that \( a = 0 \). Then from the bottom equation, it follows that \( b = 0 \) also. Thus the three functions are linearly independent.

The main theorem is the following theorem, called the replacement or exchange theorem. It uses the argument of the second half of the above proposition repeatedly.

**Theorem 3.1.5** Let \( \{u_1, \ldots, u_r\}, \{v_1, \ldots, v_s\} \) be subsets of a vector space \( V \) with field of scalars \( \mathbb{F} \) and suppose each \( u_i \in \text{span} (v_1, \ldots, v_s) \). Then \( r \leq s \). In words, linearly independent sets are no longer than spanning sets.

**Proof:** Say \( r > s \). By assumption, \( u_1 = \sum_i b_i v_i \). Not all of the \( b_i \) can equal 0 because if this were so, you would have \( u_1 = 0 \) which would violate the assumption that \( \{u_1, \ldots, u_r\} \) is linearly independent. You could write

\[ 1u_1 + 0u_2 + \cdots + 0u_r = 0 \]

since \( u_1 = 0 \). Thus some \( v_i \) say \( v_{i_1} \) is a linear combination of the vector \( u_1 \) along with the \( v_j \) for \( j \neq i \). It follows that the span of \( \{u_1, v_1, \ldots, \hat{v}_{i_1}, \ldots, v_s\} \) includes each of the \( u_i \) where the hat indicates that \( v_{i_1} \) has been omitted from the list of vectors. Now suppose each \( u_i \) is in

\[ \text{span} (u_1, \ldots, u_k, v_1, \ldots, \hat{v}_{i_1}, \ldots, \hat{v}_{i_k}, \ldots, v_s) \]

where the vectors \( \hat{v}_{i_1}, \ldots, \hat{v}_{i_k} \) have been omitted for \( k \leq s \). Then there are scalars \( c_i \) and \( d_i \) such that

\[ u_{k+1} = \sum_{i=1}^{k} c_i u_i + \sum_{j \notin \{i_1, \ldots, i_k\}} d_j v_j \]

By the assumption that \( \{u_1, \ldots, u_r\} \) is linearly independent, not all of the \( d_j \) can equal 0. Why? Therefore, there exists \( i_{k+1} \notin \{i_1, \ldots, i_k\} \) such that \( d_{i_{k+1}} \neq 0 \). Hence one can solve for \( v_{i_{k+1}} \) as a linear combination of \( \{u_1, \ldots, u_r\} \) and the \( v_j \) for \( j \notin \{i_1, \ldots, i_k, i_{k+1}\} \). Thus we can replace this \( v_{i_{k+1}} \) by a linear combination of these vectors, and so the \( u_j \) are in

\[ \text{span} (u_1, \ldots, u_k, u_{k+1}, v_1, \ldots, \hat{v}_{i_1}, \ldots, \hat{v}_{i_k}, \hat{v}_{i_{k+1}}, \ldots, v_s) \]
Continuing this replacement process, it follows that since \( r > s \), one can eliminate all of the vectors \( \{v_1, \cdots, v_s\} \) and obtain that the \( u_i \) are contained in span \((u_1, \cdots, u_s)\). But this is impossible because then you would have \( u_{s+1} \in \text{span}(u_1, \cdots, u_s) \) which is impossible since these vectors \( \{u_1, \cdots, u_r\} \) are linearly independent. It follows that \( r \leq s \).

Next is the definition of dimension and basis of a vector space.

**Definition 3.1.6** Let \( V \) be a vector space with field of scalars \( \mathbb{F} \). A subset \( S \) of \( V \) is a basis for \( V \) means that

1. \( \text{span}(S) = V \)
2. \( S \) is linearly independent.

The plural of basis is bases. It is this way to avoid hissing when referring to it.

The dimension of a vector space is the number of vectors in a basis. A vector space is finite dimensional if it equals the span of some finite set of vectors.

**Lemma 3.1.7** Let \( S \) be a linearly independent set of vectors in a vector space \( V \). Suppose \( v \notin \text{span}(S) \). Then \( \{S, v\} \) is also a linearly independent set of vectors.

**Proof:** Suppose \( \{u_1, \cdots, u_n, v\} \) is a finite subset of \( S \) and

\[
av + \sum_{i=1}^{n} b_i u_i = 0
\]

where \( a, b_1, \cdots, b_n \) are scalars. Does it follow that each of the \( b_i \) equals zero and that \( a = 0 \)? If so, then this shows that \( \{S, v\} \) is indeed linearly independent. First note that \( a = 0 \) since if not, you could write

\[
v = \sum_{i=1}^{n} \frac{b_i}{a} u_i
\]

contrary to the assumption that \( v \notin \text{span}(S) \). Hence you have \( a = 0 \) and also

\[
\sum_{i} b_i u_i = 0
\]

But \( S \) is linearly independent and so by assumption each \( b_i = 0 \).

**Proposition 3.1.8** Let \( V \) be a finite dimensional nonzero vector space with field of scalars \( \mathbb{F} \). Then it has a basis and also any two bases have the same number of vectors so the above definition of a basis is well defined.

**Proof:** Pick \( u_1 \neq 0 \). If \( \text{span}(u_1) = V \), then this is a basis. If not, there exists \( u_2 \notin \text{span}(u_1) \). Then by Lemma 3.1.3, \( \{u_1, u_2\} \) is linearly independent. If \( \text{span}(u_1, u_2) = V \), stop. You have a basis. Otherwise, there exists \( u_3 \notin \text{span}(u_1, u_2) \). Then by Lemma 3.1.3, \( \{u_1, u_2, u_3\} \) is linearly independent. Continue this way. Eventually the process yields \( \{u_1, \cdots, u_n\} \) which is linearly independent and \( \text{span}(u_1, \cdots, u_n) = V \). Otherwise there would exist a linearly independent set of \( k \) vectors for all \( k \). However, by assumption, there is a finite set of vectors \( \{v_1, \cdots, v_s\} \) such that \( \text{span}(v_1, \cdots, v_s) = V \). Therefore, \( k \leq s \). Thus there is a basis for \( V \).

If \( \{v_1, \cdots, v_s\}, \{u_1, \cdots, u_r\} \) are two bases, then since they both span \( V \) and are both linearly independent, it follows from Theorem 3.1.3 that \( r \leq s \) and \( s \leq r \).

As a specific example, consider \( \mathbb{F}^n \) as the vector space. As mentioned above, these are the mappings from \((1, \cdots, n)\) to the field \( \mathbb{F} \). It was shown in Example 3.1.9 that this is indeed a vector space with field of scalars \( \mathbb{F} \). We usually think of this \( \mathbb{F}^n \) as the set of ordered \( n \) tuples

\[
\{(x_1, \cdots, x_n) : x_i \in \mathbb{F}\}
\]
with addition and scalar multiplication defined as
\[(x_1, \cdots, x_n) + (\hat{x}_1, \cdots, \hat{x}_n) = (x_1 + \hat{x}_1, \cdots, x_n + \hat{x}_n)\]
\[\alpha (x_1, \cdots, x_n) = (\alpha x_1, \cdots, \alpha x_n)\]
Also, when referring to vectors in \(\mathbb{F}^n\), it is customary to denote them as bold faced letters, which is a convention I will begin to observe at this point. It is also more convenient to write these vectors in \(\mathbb{F}^n\) as columns of numbers. Thus
\[\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}\]
There is a fundamental concept known as linear independence.

**Observation 3.1.9** \(\mathbb{F}^n\) has dimension \(n\). To see this, note that a basis is \(\mathbf{e}_1, \cdots, \mathbf{e}_n\) where
\[\mathbf{e}_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}\]
the vector in \(\mathbb{F}^n\) which has a 1 in the \(i^{th}\) position and a zero everywhere else.

To see this, note that
\[\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \sum_{i=1}^{n} x_i \mathbf{e}_i\]
and that if
\[\mathbf{0} = \sum_{i=1}^{n} x_i \mathbf{e}_i\]
then
\[\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}\]
so each \(x_i\) is zero. Thus this set of vectors is a spanning set and is linearly independent so it is a basis. There are \(n\) of these vectors and so the dimension of \(\mathbb{F}^n\) is indeed \(n\).

There is a fundamental observation about linear combinations of vectors in \(\mathbb{F}^n\) which is stated next.

**Theorem 3.1.10** Let \(\mathbf{a}_1, \cdots, \mathbf{a}_n\) be vectors in \(\mathbb{F}^m\) where \(m < n\). Then there exist scalars \(x_1, \cdots, x_n\) not all equal to zero such that
\[x_1 \mathbf{a}_1 + \cdots + x_n \mathbf{a}_n = \mathbf{0}\]

**Proof:** If the conclusion were not so, then by definition, \(\{\mathbf{a}_1, \cdots, \mathbf{a}_n\}\) would be independent. However, there is a spanning set with only \(m\) vectors, namely \(\{\mathbf{e}_1, \cdots, \mathbf{e}_m\}\) contrary to Theorem 3.1.5. Since these vectors cannot be independent, they must be dependent which is the conclusion of the theorem.
3.2 Subspaces

The notion of a subspace is of great importance in applications. Here is what is meant by a subspace.

**Definition 3.2.1** Let $V$ be a vector space with field of scalars $\mathbb{F}$. Then let $W \subseteq V, W \neq \emptyset$. That is, $W$ is a non-empty subset of $V$. Then $W$ is a **subspace** of $V$ if whenever $\alpha, \beta$ are scalars and $u, v$ are vectors in $W$, it follows that $\alpha u + \beta v \in W$.

In words, $W$ is closed with respect to linear combinations.

The fundamental result about subspaces is that they are themselves vector spaces.

**Theorem 3.2.2** Let $W$ be a non-zero subset of $V$ a vector space with field of scalars $\mathbb{F}$. Then it is a subspace if and only if is itself a vector space with field of scalars $\mathbb{F}$.

**Proof:** Suppose $W$ is a subspace. Why is it a vector space? To be a vector space, the operations of addition and scalar multiplication must satisfy the axioms for a vector space. However, all of these are obvious because it is a subset of $V$. The only thing which is not obvious is whether $0$ is in $W$ and whether $-u \in W$ whenever $u$ is. But these follow right away from Proposition 3.1.2 because if $u \in W, (-1) u = -u \in W$ by the fact that $W$ is closed with respect to linear combinations, in particular multiplication by the scalar $-1$. Similarly, take $u \in W$. Then $0 = 0u \in W$. As to $+$ being an operation on $W$, this also follows because for $u, v \in W, u + v \in W$. Thus if it is a subspace, it is indeed a vector space.

Conversely, suppose it is a vector space. Then by definition, it is closed with respect to linear combinations and so it is a subspace.

This leads to the following simple result.

**Proposition 3.2.3** Let $W$ be a nonzero subspace of a finite dimensional vector space $V$ with field of scalars $\mathbb{F}$. Then $W$ is also a finite dimensional vector space.

**Proof:** Suppose span $(v_1, \cdots, v_n) = V$. Using the same construction of Proposition 3.1.8, the same process must stop after $k \leq n$ steps since otherwise one could obtain a linearly independent set of vectors with more vectors in it than a spanning set. Thus it has a basis with no more than $n$ vectors.

**Example 3.2.4** Show that $W = \{(x, y, z) \in \mathbb{R}^3 : x - 2y - z = 0\}$ is a subspace of $\mathbb{R}^3$. Find a basis for it.

You have from the equation that $x = 2y + z$ and so any vector in this set is of the form

$$
\begin{pmatrix}
2y + z \\
y \\
z
\end{pmatrix} : y, z \in \mathbb{R}
$$

Conversely, any vector which is of the above form satisfies the condition to be in $W$. Therefore, $W$ is of the form

$$
y \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}
$$

where $y, z$ are scalars. Hence it equals the span of the two vectors in $\mathbb{R}^3$ in the above. Are the two vectors linearly independent? If so, they will be a basis. Suppose then that

$$
y \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
$$
Then from the second position, $y = 0$. It follows then that $z = 0$ also and so the two vectors form a linearly independent set. Hence a basis for $W$ is

$$\left\{ \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix} \right\}$$

The dimension of this subspace is also 2.

**Example 3.2.5** Show that

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \\ 4 \end{pmatrix}$$

is a basis for $\mathbb{R}^3$.

There are two things to show, that the set of vectors is independent and that it spans $\mathbb{R}^3$. Thus we need to verify that there is exactly one solution to the system of equations

$$x \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + y \begin{pmatrix} 1 \\ 3 \\ 3 \end{pmatrix} + z \begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

for any choice of the right side. Recall how to do this. You set up the augmented matrix and then row reduce it.

$$\begin{pmatrix} 1 & 1 & 0 & a \\ 1 & 3 & 1 & b \\ 1 & 3 & 4 & c \end{pmatrix}$$

After some row operations, this yields

$$\begin{pmatrix} 1 & 0 & \frac{3}{2}a - \frac{2}{3}b + \frac{1}{6}c \\ 0 & 1 & \frac{3}{2}b - \frac{1}{2}a - \frac{1}{6}c \\ 0 & 0 & \frac{1}{3}c - \frac{1}{3}b \end{pmatrix}$$

Thus there is a unique solution to the system of equations. This shows that the set of vectors is a basis because one solution when the right side of the system equals the zero vector equals $x = y = z = 0$. Therefore, from what was just done, it is the only solution and so the vectors are linearly independent. As to the span of the vectors equalling $\mathbb{R}^3$, this was just shown also.

**Example 3.2.6** Show that

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ -4 \end{pmatrix}$$

is not a basis for $\mathbb{R}^3$.

You can do it the same way. It is really a question about whether there exists a unique solution to the system

$$x \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + y \begin{pmatrix} 1 \\ 3 \\ 3 \end{pmatrix} + z \begin{pmatrix} 1 \\ 1 \\ -4 \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$
for any choice of the right side. The augmented matrix is
\[
\begin{pmatrix}
1 & 1 & 1 & a \\
1 & 1 & 1 & b \\
1 & 3 & -4 & c
\end{pmatrix}
\]
After row reduction, this yields
\[
\begin{pmatrix}
1 & 1 & 1 & a \\
0 & 2 & -5 & c-a \\
0 & 0 & 0 & b-a
\end{pmatrix}
\]
Thus there is no solution to the equation unless \( b = a \). It follows the span of the given vectors is not all of \( \mathbb{R}^3 \) and so this cannot be a basis.

**Example 3.2.7** Show that
\[
\begin{pmatrix}
1 \\
1 \\
3
\end{pmatrix}, \begin{pmatrix}
1 \\
1 \\
1
\end{pmatrix}
\]
is not a basis for \( \mathbb{R}^3 \).

If the span of these vectors were all of \( \mathbb{R}^3 \), this would contradict Theorem 3.1.5 because it would be a spanning set which is shorter than a linearly independent set \( \{e_1, e_2, e_3\} \).

**Example 3.2.8** Show that
\[
\begin{pmatrix}
1 \\
1 \\
1
\end{pmatrix}, \begin{pmatrix}
1 \\
1 \\
3
\end{pmatrix}, \begin{pmatrix}
1 \\
0 \\
0
\end{pmatrix}, \begin{pmatrix}
1 \\
1 \\
1
\end{pmatrix}
\]
is not a basis for \( \mathbb{R}^3 \).

If it were a basis, then it would need to be linearly independent but this cannot happen because it would contradict Theorem 3.1.5 by being an independent set of vectors which is longer than a spanning set.

**Theorem 3.2.9** If \( V \) is an \( n \) dimensional vector space and if \( \{u_1, \ldots, u_n\} \) is a linearly independent set, then it is a basis. If \( m > n \) then \( \{v_1, \ldots, v_m\} \) is a dependent set. If \( V = \text{span} (w_1, \ldots, w_m) \), then \( m \geq n \) and there is a subset \( \{u_1, \ldots, u_n\} \subseteq \{w_1, \ldots, w_m\} \) such that \( \{u_1, \ldots, u_n\} \) is a basis. If \( \{u_1, \ldots, u_k\} \) is linearly independent, then there exists \( \{u_1, \ldots, u_k, \ldots, u_n\} \) which is a basis.

**Proof:** Say \( \{u_1, \ldots, u_n\} \) is linearly independent. Is \( \text{span} (u_1, \ldots, u_n) = V \)? If not, there would be \( w \notin \text{span} (u_1, \ldots, u_n) \) and then by Lemma 3.1.7 \( \{u_1, \ldots, u_n, w\} \) would be linearly independent which contradicts Theorem 3.1.5. As to the second claim, \( \{v_1, \ldots, v_m\} \) cannot be linearly independent because this would contradict Theorem 3.1.5 and so it is dependent.

Now say \( V = \text{span} (w_1, \ldots, w_m) \). By Theorem 3.1.5 again, you must have \( m \geq n \) since spanning sets are at least as long as linearly independent sets, one of which is a basis having \( n \) vectors. If \( w_1 \) is in the span of the other vectors, delete it. Then consider \( w_2 \). If it is in the span of the other vectors, delete it. Continue this way till a shorter list is obtained with the property that no vector is a linear combination of the others, but its span is still \( V \). By Proposition 3.1.8, the resulting list of vectors is linearly independent and is therefore, a basis since it spans \( V \).

Now suppose for \( k < n \), \( \{u_1, \ldots, u_k\} \) is linearly independent. Follow the process of Proposition 3.1.8, adding in vectors not in the span and obtaining successively larger linearly independent sets till the process ends. The resulting list must be a basis. □
3.3 Exercises

1. Show that the following are subspaces of the set of all functions defined on \([a, b]\).
   
   (a) polynomials of degree \(\leq n\)
   (b) polynomials
   (c) continuous functions
   (d) differentiable functions

2. Show that every subspace of a finite dimensional vector space \(V\) is the span of some vectors. It was done above but go over it in your own words.

3. In \(\mathbb{R}^2\) define a funny addition by \((x, y) + (\hat{x}, \hat{y}) \equiv (3x + 3\hat{x}, y + \hat{y})\) and let scalar multiplication be the usual thing. Would this be a vector space with these operations?

4. Determine which of the following are subspaces of \(\mathbb{R}^m\) for some \(m\). \(a, b\) are just given numbers in what follows.
   
   (a) \(\{(x, y) \in \mathbb{R}^2 : ax + by = 0\}\)
   (b) \(\{(x, y) \in \mathbb{R}^2 : ax + by \geq y\}\)
   (c) \(\{(x, y) \in \mathbb{R}^2 : ax + by = 1\}\)
   (d) \(\{(x, y) \in \mathbb{R}^2 : xy = 0\}\)
   (e) \(\{(x, y) \in \mathbb{R}^2 : y \geq 0\}\)
   (f) \(\{(x, y) \in \mathbb{R}^2 : x > 0 \text{ or } y > 0\}\)
   (g) For those who recall the cross product, \(\{x \in \mathbb{R}^3 : a \times x = 0\}\)
   (h) For those who recall the dot product, \(\{x \in \mathbb{R}^m : x \cdot a = 0\}\)
   (i) \(\{x \in \mathbb{R}^n : x \cdot a \geq 0\}\)
   (j) \(\{x \in \mathbb{R}^m : x \cdot s = 0\} \text{ for all } s \in S, S \neq \emptyset, S \subseteq \mathbb{R}^m\}. \text{This is known as } S^\perp.

5. Show that \(\{(x, y, z) \in \mathbb{R}^3 : x + y - z = 0\}\) is a subspace and find a basis for it.

6. In the subspace of polynomials on \([0, 1]\), show that the vectors \(\{1, x, x^2, x^3\}\) are linearly independent. Show these vectors are a basis for the vector space of polynomials of degree no more than 3.

7. Determine whether the real valued functions defined on \(\mathbb{R}\)

\[\{x^2 + 1, x^3 + 2x^2 + x, x^3 + 2x^2 - 1, x^3 + x^2 + x\}\]

are linearly independent. Is this a basis for the subspace of polynomials of degree no more than 3? Explain why or why not.

8. Determine whether the real valued functions defined on \(\mathbb{R}\)

\[\{x^2 + 1, x^3 + 2x^2 + x, x^3 + 2x^2 + x, x^3 + x^2 + x\}\]

are linearly independent. Is this a basis for the subspace of polynomials of degree no more than 3? Explain why or why not.
9. Show that the following are each a basis for \( \mathbb{R}^3 \).

   (a) \[
   \begin{pmatrix} 3 \\ 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 4 \end{pmatrix}
   \]

   (b) \[
   \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} -2 \\ 3 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 6 \end{pmatrix}
   \]

   (c) \[
   \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}
   \]

   (d) \[
   \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 6 \end{pmatrix}
   \]

10. Show that each of the following is not a basis for \( \mathbb{R}^3 \). Explain why they fail to be a basis.

   (a) \[
   \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 5 \end{pmatrix}
   \]

   (b) \[
   \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 5 \end{pmatrix}
   \]

   (c) \[
   \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}
   \]

   (d) \[
   \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix}
   \]

   (e) \[
   \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}
   \]

11. Suppose \( B \) is a subset of the set of complex valued functions, none equal to 0 and defined on \( \Omega \) and it has the property that if \( f, g \) are different, then \( fg = 0 \). Show that \( B \) must be linearly independent.

12. Suppose you have continuous real valued functions defined on \([0, 1] \), \( \{f_1, f_2, \ldots, f_n\} \) and these satisfy

\[
\int_0^1 f_i(x) f_j(x) \, dx = \delta_{ij} \equiv \begin{cases} 
1 & \text{if } i = j \\ 
0 & \text{if } i \neq j
\end{cases}
\]

   Show that these functions must be linearly independent.

13. Show that the real valued functions \( \cos(2x), 1, \cos^2(x) \) are linearly dependent.

14. Show that the real valued functions \( e^x \sin(2x), e^x \cos(2x) \) are linearly independent.
Let the field of scalars be \( \mathbb{Q} \) and let the vector space be all vectors (real numbers) of the form \( a + b\sqrt{2} \) for \( a, b \in \mathbb{Q} \). Show that this really is a vector space and find a basis for it.

Consider the two vectors \( \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix} \) in \( \mathbb{R}^2 \). Show that these are linearly independent.

Now consider \( \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix} \) in \( \mathbb{Z}_2 \) where the numbers are interpreted as residue classes. Are these vectors linearly independent? If not, give a nontrivial linear combination which is 0.

Is \( \mathbb{C} \) a vector space with field of scalars \( \mathbb{R} \)? If so, what is the dimension of this vector space? Give a basis.

Is \( \mathbb{C} \) a vector space with field of scalars \( \mathbb{C} \)? If so, what is the dimension? Give a basis.

The space of real valued continuous functions on \([0,1]\) usually denoted as \( C([0,1]) \) is a vector space with field of scalars \( \mathbb{R} \). Explain why it is not a finite dimensional vector space.

Suppose two vector spaces \( V, W \) have the same field of scalars \( \mathbb{F} \). Show that \( V \cap W \) is a subspace of both \( V \) and \( W \).

If \( V, W \) are two sub spaces of a vector space \( U \), define \( V + W \equiv \{ v + w : v \in V, w \in W \} \). Show that this is a subspace of \( U \).

If \( V, W \) are two sub spaces of a vector space \( U \), consider \( V \cup W \), the vectors which are in either \( V \) or \( W \). Will this be a subspace of \( U \)? If so, prove it is the case and if not, give an example which shows that it is not necessarily true.

Let \( V, W \) be vector spaces. A function \( T : V \to W \) is called a linear transformation if whenever \( \alpha, \beta \) are scalars and \( u, v \) are vectors in \( V \), it follows that

\[
T(\alpha u + \beta v) = \alpha Tu + \beta Tv.
\]

Then \( \ker(T) \equiv \{ u \in V : Tu = 0 \} \), \( \text{Im}(T) \equiv \{ Tu : u \in V \} \). Show the first of these is a subspace of \( V \) and the second is a subspace of \( W \).

In the situation of the above problem, where \( T \) is a linear transformation, suppose \( S \) is a linearly independent subset of \( W \). Define \( T^{-1}(S) \equiv \{ u \in V : Tu \in S \} \). Show that \( T^{-1}(S) \) is linearly independent.

In the situation of the above problems, rank \( (T) \) is defined as the dimension of \( \text{Im}(T) \). Also the nullity of \( T \), denoted as null \( (T) \) is defined as the dimension of \( \ker(T) \). In this problem, you will show that if the dimension of \( V \) is \( n \), then rank \( (T) + \text{null}(T) = n \).

(a) Let a basis for \( \ker(T) \) be \( \{ z_1, \ldots, z_r \} \). Let a basis for \( \text{Im}(T) \) be \( \{ Tv_1, \ldots, Tv_s \} \). You need to show that \( r + s = n \). Begin with \( u \in V \) and consider \( Tu \). It is a linear combination of \( \{ Tv_1, \ldots, Tv_s \} \) say \( \sum_{i=1}^s a_iTv_i \). Why?

(b) Next explain why \( T(u - \sum_{i=1}^s a_i v_i) = 0 \). Then explain why there are scalars \( b_j \) such that \( u - \sum_{i=1}^s a_i v_i = \sum_{j=1}^r b_j z_j \).

(c) Observe that \( V = \text{span}\{z_1, \ldots, z_r, v_1, \ldots, v_s\} \). Why?

(d) Finally show that \( \{ z_1, \ldots, z_r, v_1, \ldots, v_s \} \) is linearly independent. Thus \( n = r + s \).
3.4 Polynomials And Fields

As an application of the theory of vector spaces, this section considers the problem of field extensions. When you have a polynomial like $x^2 - 3$ which has no rational roots, it turns out you can enlarge the field of rational numbers to obtain a larger field such that this polynomial does have roots in this larger field. I am going to discuss a systematic way to do this. It will turn out that for any polynomial with coefficients in any field, there always exists a possibly larger field such that the polynomial has roots in this larger field. This book mainly features the field of real or complex numbers but this procedure will show how to obtain many other fields. The ideas used in this development are the same as those used later in the material on linear transformations but slightly easier.

Here is an important idea concerning equivalence relations which I hope is familiar.

**Definition 3.4.1** Let $S$ be a set. The symbol, $\sim$, is called an equivalence relation on $S$ if it satisfies the following axioms.

1. $x \sim x$ for all $x \in S$. (Reflexive)
2. If $x \sim y$ then $y \sim x$. (Symmetric)
3. If $x \sim y$ and $y \sim z$, then $x \sim z$. (Transitive)

**Definition 3.4.2** $[x]$ denotes the set of all elements of $S$ which are equivalent to $x$ and $[x]$ is called the equivalence class determined by $x$ or just the equivalence class of $x$.

Also recall the notion of equivalence classes.

**Theorem 3.4.3** Let $\sim$ be an equivalence class defined on a set, $S$ and let $\mathcal{H}$ denote the set of equivalence classes. Then if $[x]$ and $[y]$ are two of these equivalence classes, either $x \sim y$ and $[x] = [y]$ or it is not true that $x \sim y$ and $[x] \cap [y] = \emptyset$.

**Definition 3.4.4** Let $\mathbb{F}$ be a field, for example the rational numbers, and denote by $\mathbb{F}[x]$ the polynomials having coefficients in $\mathbb{F}$. Suppose $p(x)$ is a polynomial. Let $a(x) \sim b(x)$ (a polynomial is similar to $b(x)$) when

$$a(x) - b(x) = k(x)p(x)$$

for some polynomial $k(x)$.

**Proposition 3.4.5** In the above definition, $\sim$ is an equivalence relation.

**Proof:** First of all, note that $a(x) \sim a(x)$ because their difference equals $0p(x)$. If $a(x) \sim b(x)$, then $a(x) - b(x) = k(x)p(x)$ for some $k(x)$. But then $b(x) - a(x) = -k(x)p(x)$ and so $b(x) \sim a(x)$.

Next suppose $a(x) \sim b(x)$ and $b(x) \sim c(x)$. Then $a(x) - b(x) = k(x)p(x)$ for some polynomial $k(x)$ and also $b(x) - c(x) = l(x)p(x)$ for some polynomial $l(x)$. Then

$$a(x) - c(x) = a(x) - b(x) + b(x) - c(x) = k(x)p(x) + l(x)p(x) = (l(x) + k(x))p(x)$$

and so $a(x) \sim c(x)$ and this shows the transitive law. ■

With this proposition, here is another definition which essentially describes the elements of the new field. It will eventually be necessary to assume that the polynomial $p(x)$ in the above definition is irreducible so I will begin assuming this.

**Definition 3.4.6** Let $\mathbb{F}$ be a field and let $p(x) \in \mathbb{F}[x]$ be a monic irreducible polynomial of degree greater than 0. Thus there is no polynomial having coefficients in $\mathbb{F}$ which divides $p(x)$ except for itself and constants, and its leading coefficient is 1. For the similarity relation defined in Definition
In the situation of Definition\ref{eq:1.4.7}, the definitions of addition and multiplication are well defined.

\[ [a(x)] + [b(x)] \equiv [a(x) + b(x)] \]
\[ [a(x)] [b(x)] \equiv [a(x) b(x)] \]

This collection of equivalence classes is sometimes denoted by \( F[x]/(p(x)) \).

**Proposition 3.4.7** In the situation of Definition\ref{eq:1.4.7} where \( p(x) \) is a monic irreducible polynomial, the following are valid.

1. \( p(x) \) and \( q(x) \) are relatively prime for any \( q(x) \in F[x] \) which is not a multiple of \( p(x) \).
2. The definitions of addition and multiplication are well defined.
3. If \( a, b \in F \) and \([a] = [b] \), then \( a = b \). Thus \( F \) can be considered a subset of \( F[x]/(p(x)) \).
4. \( F[x]/(p(x)) \) is a field in which the polynomial \( p(x) \) has a root.
5. \( F[x]/(p(x)) \) is a vector space with field of scalars \( F \) and its dimension is \( m \) where \( m \) is the degree of the irreducible polynomial \( p(x) \).

**Proof:** First consider the claim about \( p(x) \), \( q(x) \) being relatively prime. If \( \psi(x) \) is the greatest common divisor, (the monic polynomial of largest degree which divides both) it follows \( \psi(x) \) is either equal to \( p(x) \) or 1. If it is \( p(x) \), then \( q(x) \) is a multiple of \( p(x) \) which does not happen. If it is 1, then by definition, the two polynomials are relatively prime.

To show the operations are well defined, suppose

\[ [a(x)] = [a'(x)], [b(x)] = [b'(x)] \]

It is necessary to show

\[ [a(x) + b(x)] = [a'(x) + b'(x)] \]
\[ [a(x) [b(x)] = [a'(x) b'(x)] \]

Consider the second of the two.

\[
\begin{align*}
    a'(x) b'(x) - a(x) b(x) \\
    = a'(x) b'(x) - a(x) b'(x) + a(x) b'(x) - a(x) b(x) \\
    = b'(x)(a'(x) - a(x)) + a(x)(b'(x) - b(x))
\end{align*}
\]

Now by assumption \( (a'(x) - a(x)) \) is a multiple of \( p(x) \) as is \( (b'(x) - b(x)) \), so the above is a multiple of \( p(x) \) and by definition this shows \([a(x) b(x)] = [a'(x) b'(x)] \). The case for addition is similar.

Now suppose \([a] = [b] \). This means \( a - b = k(x)p(x) \) for some polynomial \( k(x) \). Then \( k(x) \) must equal 0 since otherwise the two polynomials \( a - b \) and \( k(x)p(x) \) could not be equal because they would have different degree.

It is clear that the axioms of a field are satisfied except for the one which says that non zero elements of the field have a multiplicative inverse. Let \([q(x)] \in F[x]/(p(x)) \) where \([q(x)] \neq [0] \). Then \( q(x) \) is not a multiple of \( p(x) \) and so by the first part, \( q(x), p(x) \) are relatively prime. Thus there exist \( n(x), m(x) \) such that

\[ 1 = n(x) q(x) + m(x) p(x) \]

Hence

\[ [1] = [1 - n(x) p(x)] = [n(x) q(x)] = [n(x)] [q(x)] \]

which shows that \([q(x)]^{-1} = [n(x)] \). Thus this is a field. The polynomial has a root in this field because if

\[ p(x) = x^m + a_m x^{m-1} + \cdots + a_1 x + a_0, \]
Thus \([x]\) is a root of this polynomial in the field \(F[x]/(p(x))\).

Consider the last claim. Let \(f(x) \in F[x]/(p(x))\). Thus \([f(x)]\) is a typical thing in \(F[x]/(p(x))\). Then from the division algorithm,

\[
f(x) = p(x)q(x) + r(x)
\]

where \(r(x)\) is either 0 or has degree less than the degree of \(p(x)\). Thus

\[
[r(x)] = [f(x) - p(x)q(x)] = [f(x)]
\]

but clearly \([r(x)] \in \text{span}(\{[1], \ldots, [x]^{m-1}\})\). Thus \(\text{span}(\{[1], \ldots, [x]^{m-1}\}) = F[x]/(p(x))\). Then \(\{[1], \ldots, [x]^{m-1}\}\) is a basis if these vectors are linearly independent. Suppose then that

\[
\sum_{i=0}^{m-1} c_i [x]^i = \left[ \sum_{i=0}^{m-1} c_i x^i \right] = 0
\]

Then you would need to have \(p(x)/\sum_{i=0}^{m-1} c_i x^i\) which is impossible unless each \(c_i = 0\) because \(p(x)\) has degree \(m\).

From the above theorem, it makes perfect sense to write \(b\) rather than \([b]\) if \(b \in F\). Then with this convention,

\[
[b\phi(x)] = [b]\phi(x) = b\phi(x).
\]

This shows how to enlarge a field to get a new one in which the polynomial has a root. By using a succession of such enlargements, called field extensions, there will exist a field in which the given polynomial can be factored into a product of polynomials having degree one. The field you obtain in this process of enlarging in which the given polynomial factors in terms of linear factors is called a splitting field.

**Remark 3.4.8** The polynomials consisting of all polynomial multiples of \(p(x)\), denoted by \((p(x))\), is called an ideal. An ideal \(I\) is a subset of the commutative ring (Here the ring is \(F[x]\)) with unity consisting of all polynomials which is itself a ring and which has the property that whenever \(f(x) \in F[x]\), and \(g(x) \in I\), \(f(x)g(x) \in I\). In this case, you could argue that \((p(x))\) is an ideal and that the only ideal containing it is itself or the entire ring \(F[x]\). This is called a maximal ideal.

**Example 3.4.9** The polynomial \(x^2 - 2\) is irreducible in \(\mathbb{Q}[x]\). This is because if \(x^2 - 2 = p(x)q(x)\) where \(p(x), q(x)\) both have degree less than 2, then they both have degree 1. Hence you would have \(x^2 - 2 = (x+a)(x+b)\) which requires that \(a+b = 0\) so this factorization is of the form \((x-a)(x+a)\) and now you need to have \(a = \sqrt{2} \notin \mathbb{Q}\). Now \(\mathbb{Q}[x]/(x^2-2)\) is of the form \(a+b[x]\) where \(a, b \in \mathbb{Q}\) and \([x]^2 - 2 = 0\). Thus one can regard \([x]\) as \(\sqrt{2}\). \(\mathbb{Q}[x]/(x^2-2)\) is of the form \(a+b\sqrt{2}\).

In the above example, \([x^2 + x]\) is not zero because it is not a multiple of \(x^2 - 2\). What is \([x^2 + x]^{-1}\)? You know that the two polynomials are relatively prime and so there exists \(n(x), m(x)\) such that

\[
1 = n(x)(x^2-2) + m(x)(x^2+x)
\]

Thus \([m(x)] = [x^2 + x]^{-1}\). How could you find these polynomials? First of all, it suffices to consider only \(n(x)\) and \(m(x)\) having degree less than 2.

\[
1 = (ax+b)(x^2-2)+(cx+d)(x^2+x)
\]

\[
1 = ax^3-2b+bx^2+cx^2+cx^3+dx^2-2ax+dx
\]
Now you solve the resulting system of equations.

\[
a = \frac{1}{2}, \quad b = -\frac{1}{2}, \quad c = -\frac{1}{2}, \quad d = 1
\]

Then the desired inverse is \([-\frac{1}{2}x + 1]\). To check,

\[
\left(-\frac{1}{2}x + 1\right)(x^2 + x) - 1 = -\frac{1}{2}(x - 1)(x^2 - 2)
\]

Thus \([-\frac{1}{2}x + 1]\) \([x^2 + x] - [1] = [0]\).

The above is an example of something general described in the following definition.

**Definition 3.4.10** Let \(F \subseteq K\) be two fields. Then clearly \(K\) is also a vector space over \(F\). Then also, \(K\) is called a finite field extension of \(F\) if the dimension of this vector space, denoted by \([K : F]\) is finite.

There are some easy things to observe about this.

**Proposition 3.4.11** Let \(F \subseteq K \subseteq L\) be fields. Then \([L : F] = [L : K][K : F]\).

**Proof:** Let \(\{l_i\}_{i=1}^n\) be a basis for \(L\) over \(K\) and let \(\{k_j\}_{j=1}^m\) be a basis of \(K\) over \(F\). Then if \(l \in L\), there exist unique scalars \(x_i\) in \(K\) such that

\[
l = \sum_{i=1}^n x_i l_i
\]

Now \(x_i \in K\) so there exist \(f_{ji}\) such that

\[
x_i = \sum_{j=1}^m f_{ji} k_j
\]

Then it follows that

\[
l = \sum_{i=1}^n \sum_{j=1}^m f_{ji} k_j l_i
\]

It follows that \(\{k_j l_i\}\) is a spanning set. If

\[
\sum_{i=1}^n \sum_{j=1}^m f_{ji} k_j l_i = 0
\]

Then, since the \(l_i\) are independent, it follows that

\[
\sum_{j=1}^m f_{ji} k_j = 0
\]

and since \(\{k_j\}\) is independent, each \(f_{ji} = 0\) for each \(j\) for a given arbitrary \(i\). Therefore, \(\{k_j l_i\}\) is a basis. \(\square\)

You will see almost exactly the same argument in exhibiting a basis for \(\mathcal{L}(V, W)\) the linear transformations mapping \(V\) to \(W\).

Note that if \(p(x)\) were not irreducible, then you could find a field extension \(G\) containing a root of \(p(x)\) such that \([G : F] \leq n\). You could do this by working with an irreducible factor of \(p(x)\).

**Theorem 3.4.12** Let \(p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0\) be a polynomial with coefficients in a field of scalars \(F\). There exists a larger field \(G\) and \(\{z_1, \ldots, z_n\}\) contained in \(G\), listed according to multiplicity, such that

\[
p(x) = \prod_{i=1}^n (x - z_i)
\]

This larger field is called a splitting field. Furthermore,

\([G : F] \leq n!\)
3.4.7 In the situation of the above example, find a larger field $G$ such that $p(x)$ has a root, $z_1 = [x]$, if $p$ is irreducible. Then by the Euclidean algorithm

$$p(x) = (x - z_1) q_1(x) + r$$

where $r \in F_1$. Since $p(z_1) = 0$, this requires $r = 0$. Now do the same for $q_1(x)$ that was done for $p(x)$, enlarging the field to $F_2$ if necessary, such that in this new field

$$q_1(x) = (x - z_2) q_2(x),$$

and so

$$p(x) = (x - z_1)(x - z_2) q_2(x)$$

After $n$ such extensions, you will have obtained the necessary field $G$.

Finally consider the claim about dimension. By Proposition 3.4.14, there is a larger field $G_1$ such that $p(x)$ has a root $a_1$ in $G_1$ and $[G_1 : F] \leq n$. Then

$$p(x) = (x - a_1) q_1(x)$$

Continue this way until the polynomial equals the product of linear factors. Then by Proposition 3.4.11 applied multiple times, $[G : F] \leq n!$.

**Example 3.4.13** The polynomial $x^2 + 1$ is irreducible in $\mathbb{R}[x]$, polynomials having real coefficients. To see this is the case, suppose $\psi(x)$ divides $x^2 + 1$. Then

$$x^2 + 1 = \psi(x) q(x)$$

If the degree of $\psi(x)$ is less than 2, then it must be either a constant or of the form $ax + b$. In the latter case, $-b/a$ must be a zero of the right side, hence of the left but $x^2 + 1$ has no real zeros. Therefore, the degree of $\psi(x)$ must be two and $q(x)$ must be a constant. Thus the only polynomial which divides $x^2 + 1$ are constants and multiples of $x^2 + 1$. Therefore, this shows $x^2 + 1$ is irreducible. Find the inverse of $[x^2 + x + 1]$ in the space of equivalence classes, $\mathbb{R}/(x^2 + 1)$.

You can solve this with partial fractions.

$$\frac{1}{(x^2 + 1)(x^2 + x + 1)} = \frac{x}{x^2 + 1} + \frac{x + 1}{x^2 + x + 1}$$

and so

$$1 = (-x)(x^2 + x + 1) + (x + 1)(x^2 + 1)$$

which implies

$$1 \sim (-x)(x^2 + x + 1)$$

and so the inverse is $[-x]$.

The following proposition is interesting. It was essentially proved above but to emphasize it, here it is again.

**Proposition 3.4.14** Suppose $p(x) \in \mathbb{F}[x]$ is irreducible and has degree $n$. Then every element of $G = \mathbb{F}[x]/(p(x))$ is of the form $[0]$ or $[r(x)]$ where the degree of $r(x)$ is less than $n$.

**Proof:** This follows right away from the Euclidean algorithm for polynomials. If $k(x)$ has degree larger than $n - 1$, then

$$k(x) = q(x)p(x) + r(x)$$

where $r(x)$ is either equal to 0 or has degree less than $n$. Hence

$$[k(x)] = [r(x)].$$

**Example 3.4.15** In the situation of the above example, find $[ax + b]^{-1}$ assuming $a^2 + b^2 \neq 0$. Note this includes all cases of interest thanks to the above proposition.
You can do it with partial fractions as above.

\[
\frac{1}{(x^2 + 1)(ax + b)} = \frac{b - ax}{(a^2 + b^2)(x^2 + 1)} + \frac{a^2}{(a^2 + b^2)(ax + b)}
\]

and so

\[
1 = \frac{1}{a^2 + b^2} (b - ax) (ax + b) + \frac{a^2}{(a^2 + b^2)} (x^2 + 1)
\]

Thus

\[
\frac{1}{a^2 + b^2} (b - ax) (ax + b) \sim 1
\]

and so

\[
(ax + b)^{-1} = \frac{[(b - ax)]}{a^2 + b^2} = \frac{b - a[x]}{a^2 + b^2}
\]

You might find it interesting to recall that \((ai + b)^{-1} = \frac{b - ai}{a^2 + b^2}\). Didn’t we just produce \(i\) algebraically?

### 3.4.1 The Algebraic Numbers

Each polynomial having coefficients in a field \(F\) has a splitting field. Consider the case of all polynomials \(p(x)\) having coefficients in a field \(F \subseteq \mathbb{G}\) and consider all roots which are also in \(\mathbb{G}\). The theory of vector spaces is very useful in the study of these algebraic numbers. Here is a definition.

**Definition 3.4.16** The algebraic numbers \(\mathbb{A}\) are those numbers which are in \(\mathbb{G}\) and also roots of some polynomial \(p(x)\) having coefficients in \(F\). The minimum polynomial\(^1\) of \(a \in \mathbb{A}\) is defined to be the monic polynomial \(p(x)\) having smallest degree such that \(p(a) = 0\).

The next theorem is on the uniqueness of the minimum polynomial.

**Theorem 3.4.17** Let \(a \in \mathbb{A}\). Then there exists a unique monic irreducible polynomial \(p(x)\) having coefficients in \(F\) such that \(p(a) = 0\). This polynomial is the minimum polynomial.

**Proof:** Let \(p(x)\) be a monic polynomial having smallest degree such that \(p(a) = 0\). Then \(p(x)\) is irreducible because if not, there would exist a polynomial having smaller degree which has \(a\) as a root. Now suppose \(q(x)\) is monic with smallest degree such that \(q(a) = 0\). Then

\[
q(x) = p(x)l(x) + r(x)
\]

where if \(r(x) \neq 0\), then it has smaller degree than \(p(x)\). But in this case, the equation implies \(r(a) = 0\) which contradicts the choice of \(p(x)\). Hence \(r(x) = 0\) and so, since \(q(x)\) has smallest degree, \(l(x) = 1\) showing that \(p(x) = q(x)\).

**Definition 3.4.18** For an algebraic number, let \(\deg(a)\) denote the degree of the minimum polynomial of \(a\).

Also, here is another definition.

**Definition 3.4.19** Let \(a_1, \ldots, a_m\) be in \(\mathbb{A}\). A polynomial in \(\{a_1, \ldots, a_m\}\) will be an expression of the form

\[
\sum_{k_1 \cdots k_n} a_{k_1 \cdots k_n} a_1^{k_1} \cdots a_m^{k_n}
\]

where the \(a_{k_1 \cdots k_n}\) are in \(F\), each \(k_j\) is a nonnegative integer, and all but finitely many of the \(a_{k_1 \cdots k_n}\) equal zero. The collection of such polynomials will be denoted by \(F[a_1, \ldots, a_m]\).

---

\(^1\) I grew up calling this and similar things the minimal polynomial, but I think it is better to call it the minimum polynomial because it is unique.
Now notice that for a an algebraic number, \( \mathbb{F} [a] \) is a vector space with field of scalars \( \mathbb{F} \). Similarly, for \( \{a_1, \ldots, a_m\} \) algebraic numbers, \( \mathbb{F} [a_1, \ldots, a_m] \) is a vector space with field of scalars \( \mathbb{F} \). The following fundamental proposition is important.

**Proposition 3.4.20** Let \( \{a_1, \ldots, a_m\} \) be algebraic numbers. Then

\[
\dim \mathbb{F} [a_1, \ldots, a_m] \leq \prod_{j=1}^{m} \deg (a_j)
\]

and for an algebraic number \( a \),

\[
\dim \mathbb{F} [a] = \deg (a)
\]

Every element of \( \mathbb{F} [a_1, \ldots, a_m] \) is in \( A \) and \( \mathbb{F} [a_1, \ldots, a_m] \) is a field.

**Proof:** Let the minimum polynomial of \( a \) be

\[ p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0. \]

If \( q(a) \in \mathbb{F} [a] \), then

\[ q(x) = p(x)l(x) + r(x) \]

where \( r(x) \) has degree less than the degree of \( p(x) \) if it is not zero. Hence \( q(a) = r(a) \). Thus \( \mathbb{F} [a] \) is spanned by

\[ \{1, a, a^2, \ldots, a^{n-1}\} \]

Since \( p(x) \) has smallest degree of all polynomials which have \( a \) as a root, the above set is also linearly independent. This proves the second claim.

Now consider the first claim. By definition, \( \mathbb{F} [a_1, \ldots, a_m] \) is obtained from all linear combinations of products of \( \{a_1^{k_1}, a_2^{k_2}, \ldots, a_n^{k_n}\} \) where the \( k_i \) are nonnegative integers. From the first part, it suffices to consider only \( k_j \leq \deg (a_j) \). Therefore, there exists a spanning set for \( \mathbb{F} [a_1, \ldots, a_m] \) which has

\[ \prod_{i=1}^{m} \deg (a_i) \]

entries. By Theorem 3.4.16 this proves the first claim.

Finally consider the last claim. Let \( g(a_1, \ldots, a_m) \) be a polynomial in \( \{a_1, \ldots, a_m\} \) in \( \mathbb{F} [a_1, \ldots, a_m] \). Since

\[ \dim \mathbb{F} [a_1, \ldots, a_m] = p \leq \prod_{j=1}^{m} \deg (a_j) < \infty, \]

it follows

\[ 1, g(a_1, \ldots, a_m), g(a_1, \ldots, a_m)^2, \ldots, g(a_1, \ldots, a_m)^p \]

are dependent. It follows \( g(a_1, \ldots, a_m) \) is the root of some polynomial having coefficients in \( \mathbb{F} \). Thus everything in \( \mathbb{F} [a_1, \ldots, a_m] \) is algebraic. Why is \( \mathbb{F} [a_1, \ldots, a_m] \) a field? Let \( g(a_1, \ldots, a_m) \) be as just mentioned. Then it has a minimum polynomial,

\[ p(x) = x^q + a_{q-1}x^{q-1} + \cdots + a_1x + a_0 \]

where the \( a_i \in \mathbb{F} \). Then \( a_0 \neq 0 \) or else the polynomial would not be minimum. Therefore,

\[ g(a_1, \ldots, a_m)(g(a_1, \ldots, a_m)^{q-1} + a_{q-1}g(a_1, \ldots, a_m)^{q-2} + \cdots + a_1) = -a_0 \]

and so the multiplicative inverse for \( g(a_1, \ldots, a_m) \) is

\[ \frac{g(a_1, \ldots, a_m)^{q-1} + a_{q-1}g(a_1, \ldots, a_m)^{q-2} + \cdots + a_1}{-a_0} \in \mathbb{F} [a_1, \ldots, a_m]. \]

The other axioms of a field are obvious. \( \blacksquare \)

Now from this proposition, it is easy to obtain the following interesting result about the algebraic numbers.
Theorem 3.4.21 The algebraic numbers \( \mathbb{A} \), those roots of polynomials in \( \mathbb{F}[x] \) which are in \( \mathbb{G} \), are a field.

**Proof:** By definition, each \( a \in \mathbb{A} \) has a minimum polynomial. Let \( a \neq 0 \) be an algebraic number and let \( p(x) \) be its minimum polynomial. Then \( p(x) \) is of the form

\[
x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0
\]

where \( a_0 \neq 0 \). Otherwise \( p(x) \) would not have minimum degree. Then plugging in \( a \) yields

\[
a \left( a^{n-1} + a_{n-1}a^{n-2} + \cdots + a_1 \right) (-1)^{n-1} = 1.
\]

and so \( a^{-1} = \left( a^{-n} + a_{n-1}a^{-n+2} + \cdots + a_1 \right) \in \mathbb{F}[a] \). By the proposition, every element of \( \mathbb{F}[a] \) is in \( \mathbb{A} \) and this shows that for every nonzero element of \( \mathbb{A} \), its inverse is also in \( \mathbb{A} \). What about products and sums of things in \( \mathbb{A} \)? Are they still in \( \mathbb{A} \)? Yes. If \( a, b \in \mathbb{A} \), then both \( a + b \) and \( ab \in \mathbb{F}[a,b] \) and from the proposition, each element of \( \mathbb{F}[a,b] \) is in \( \mathbb{A} \).

A typical example of what is of interest here is when the field \( \mathbb{F} \) of scalars is \( \mathbb{Q} \), the rational numbers and the field \( \mathbb{G} \) is \( \mathbb{R} \). However, you can certainly conceive of many other examples by considering the integers mod a prime, for example (See Proposition 3.4.21 on Page 46 for example) or any of the fields which occur as field extensions in the above.

There is a very interesting thing about \( \mathbb{F}[a_1 \cdots a_n] \) in the case where \( \mathbb{F} \) is infinite which says that there exists a single algebraic \( \gamma \) such that \( \mathbb{F}[a_1 \cdots a_n] = \mathbb{F}[\gamma] \). In other words, every field extension of this sort is a simple field extension. I found this fact in an early version of [III].

**Proposition 3.4.22** There exists \( \gamma \) such that \( \mathbb{F}[a_1 \cdots a_n] = \mathbb{F}[\gamma] \). Here each \( a_i \) is algebraic.

**Proof:** To begin with, consider \( \mathbb{F}[\alpha, \beta] \). Let \( \gamma = \alpha + \lambda \beta \). Then by Proposition 3.4.21, \( \gamma \) is an algebraic number and it is also clear

\[
\mathbb{F}[\gamma] \subseteq \mathbb{F}[\alpha, \beta]
\]

I need to show the other inclusion. This will be done for a suitable choice of \( \lambda \). To do this, it suffices to verify that both \( \alpha \) and \( \beta \) are in \( \mathbb{F}[\gamma] \).

Let the minimum polynomials of \( \alpha \) and \( \beta \) be \( f(x) \) and \( g(x) \) respectively. Let the distinct roots of \( f(x) \) and \( g(x) \) be \( \{\alpha_1, \alpha_2, \dots, \alpha_n\} \) and \( \{\beta_1, \beta_2, \dots, \beta_m\} \) respectively. These roots are in a field which contains splitting fields of both \( f(x) \) and \( g(x) \). Let \( \alpha = \alpha_1 \) and \( \beta = \beta_1 \). Now define

\[
h(x) = f(\alpha + \lambda \beta - \lambda x) \equiv f(\gamma - \lambda x)
\]

so that \( h(\beta) = f(\alpha) = 0 \). It follows \( (x - \beta) \) divides both \( h(x) \) and \( g(x) \). If \( (x - \eta) \) is a different linear factor of both \( g(x) \) and \( h(x) \) then it must be \( (x - \beta_j) \) for some \( \beta_j \) for some \( j > 1 \) because these are the only factors of \( g(x) \). Therefore, this would require

\[
0 = h(\beta_j) = f(\alpha_1 + \lambda \beta_1 - \lambda \beta_j)
\]

and so it would be the case that \( \alpha_1 + \lambda \beta_1 - \lambda \beta_j = \alpha_k \) for some \( k \). Hence

\[
\lambda = \frac{\alpha_k - \alpha_1}{\beta_1 - \beta_j}
\]

Now there are finitely many quotients of the above form and if \( \lambda \) is chosen to not be any of them, then the above cannot happen and so in this case, the only linear factor of both \( g(x) \) and \( h(x) \) will be \( (x - \beta) \). Choose such a \( \lambda \).

Let \( \phi(x) \) be the minimum polynomial of \( \beta \) with respect to the field \( \mathbb{F}[\gamma] \). Then this minimum polynomial must divide both \( h(x) \) and \( g(x) \) because \( h(\beta) = g(\beta) = 0 \). However, the only factor these two have in common is \( x - \beta \) and so \( \phi(x) = x - \beta \) which requires \( \beta \in \mathbb{F}[\gamma] \). Now also \( \alpha = \gamma - \lambda \beta \).
and so $a \in \mathbb{F}[\gamma]$ also. Therefore, both $\alpha, \beta \in \mathbb{F}[\gamma]$ which forces $\mathbb{F}[\alpha, \beta] \subseteq \mathbb{F}[\gamma]$. This proves the proposition in the case that $n = 2$. The general result follows right away by observing that

$$\mathbb{F}[a_1 \cdots a_n] = \mathbb{F}[a_1 \cdots a_{n-1}][a_n]$$

and using induction. \[\square\]

When you have a field $\mathbb{F}$, $\mathbb{F}(a)$ denotes the smallest field which contains both $\mathbb{F}$ and $a$. When $a$ is algebraic over $\mathbb{F}$, it follows that $\mathbb{F}(a) = \mathbb{F}[a]$. The latter is easier to think about because it just involves polynomials.

### 3.4.2 The Lindemann-Weierstrass Theorem And Vector Spaces

As another application of the abstract concept of vector spaces, there is an amazing theorem due to Weierstrass and Lindemann.

**Theorem 3.4.23** Suppose $a_1, \ldots, a_n$ are algebraic numbers, roots of a polynomial with rational coefficients, and suppose $\alpha_1, \ldots, \alpha_n$ are distinct algebraic numbers. Then

$$\sum_{i=1}^{n} a_i e^{\alpha_i} \neq 0$$

In other words, the $\{e^{\alpha_1}, \ldots, e^{\alpha_n}\}$ are independent as vectors with field of scalars equal to the algebraic numbers.

There is a proof of this later. It is long and hard but only depends on elementary considerations other than some algebra involving symmetric polynomials. See Theorem 10.2.5. It is presented here to illustrate how the language of linear algebra is useful in describing something which is really pretty exotic.

A number is transcendental, as opposed to algebraic, if it is not a root of a polynomial which has integer (rational) coefficients. Most numbers are this way but it is hard to verify that specific numbers are transcendental. That $\pi$ is transcendental follows from

$$e^0 + e^{i\pi} = 0.$$  

By the above theorem, this could not happen if $\pi$ were algebraic because then $i\pi$ would also be algebraic. Recall these algebraic numbers form a field and $i$ is clearly algebraic, being a root of $x^2 + 1$. This fact about $\pi$ was first proved by Lindemann in 1882 and then the general theorem above was proved by Weierstrass in 1885. This fact that $\pi$ is transcendental solved an old problem called squaring the circle which was to construct a square with the same area as a circle using a straight edge and compass. It can be shown that the fact $\pi$ is transcendental implies this problem is impossible.\footnote{Gilbert, the librettist of the Savoy operas, may have heard about this great achievement. In Princess Ida which opened in 1884 he has the following lines. “As for fashion they forswear it, so they say - so they say; and the circle - they will square it some fine day some fine day.” Of course it had been proved impossible to do this a couple of years before.}

### 3.5 Exercises

1. Let $p(x) \in \mathbb{F}[x]$ and suppose that $p(x)$ is the minimum polynomial for $a \in \mathbb{F}$. Consider a field extension of $\mathbb{F}$ called $\mathbb{G}$. Thus $a \in \mathbb{G}$ also. Show that the minimum polynomial of $a$ with coefficients in $\mathbb{G}$ must divide $p(x)$.

2. Here is a polynomial in $\mathbb{Q}[x]$  

$$x^2 + x + 3$$

Show it is irreducible in $\mathbb{Q}[x]$. Now consider $x^2 - x + 1$. Show that in $\mathbb{Q}[x] / (x^2 + x + 3)$ it follows that $[x^2 - x + 1] \neq 0$. Find its inverse in $\mathbb{Q}[x] / (x^2 + x + 3)$. 
3. Here is a polynomial in \( \mathbb{Q}[x] \)

\[ x^2 - x + 2 \]

Show it is irreducible in \( \mathbb{Q}[x] \). Now consider \( x + 2 \). Show that in \( \mathbb{Q}[x]/(x^2 - x + 2) \) it follows that \( [x + 2] \neq 0 \). Find its inverse in \( \mathbb{Q}[x]/(x^2 - x + 2) \).

4. Here is a polynomial in \( \mathbb{Z}_3[x] \)

\[ x^2 + x + 2 \]

Show it is irreducible in \( \mathbb{Z}_3[x] \). Show \( [x + 2] \) is not zero in \( \mathbb{Z}_3[x]/(x^2 + x + 2) \). Now find its inverse in \( \mathbb{Z}_3[x]/(x^2 + x + 2) \).

5. Suppose you have \( W \) a subspace of a finite dimensional vector space \( V \). Suppose also that \( \dim(W) = \dim(V) \). Tell why \( W = V \).

6. Suppose \( V \) is a vector space with field of scalars \( \mathbb{F} \). Let \( T \in \mathcal{L}(V,W) \), the space of linear transformations mapping \( V \) onto \( W \) where \( W \) is another vector space (See Problem 23 on Page 62). Define an equivalence relation on \( V \) as follows. \( v \sim w \) means \( v - w \in \ker(T) \).

Recall that \( \ker(T) \equiv \{ v : T(v) = 0 \} \). Show this is an equivalence relation. Now for \( [v] \) an equivalence class define \( T'([v]) \equiv Tv \). Show this is well defined. Also show that with the operations

\[
[v] + [w] \equiv [v + w] \\
\alpha [v] \equiv [\alpha v]
\]

this set of equivalence classes, denoted by \( V/\ker(T) \) is a vector space. Show next that \( T' : V/\ker(T) \to W \) is one to one. This new vector space, \( V/\ker(T) \) is called a quotient space. Show its dimension equals the difference between the dimension of \( V \) and the dimension of \( \ker(T) \).

7. Now suppose now that \( W = T(V) \). Then show that \( T' \) in the above is one to one and onto. Explain why \( \dim(V/\ker(T)) = \dim(T(V)) \). Now see Problem 23 on Page 62. Show that

\[ \text{rank}(T) + \text{null}(T) = \dim(V). \]

8. Let \( V \) be an \( n \)-dimensional vector space and let \( W \) be a subspace. Generalize the Problem 1 to define and give properties of \( V/W \). What is its dimension? What is a basis?

9. A number is transcendental if it is not the root of any nonzero polynomial with rational coefficients. As mentioned, there are many known transcendental numbers. Suppose \( \alpha \) is a real transcendental number. Show that \( \{ 1, \alpha, \alpha^2, \ldots \} \) is a linearly independent set of real numbers if the field of scalars is the rational numbers.

10. Suppose \( \mathbb{F} \) is a countable field and let \( \mathbb{A} \) be the algebraic numbers, those numbers in \( \mathbb{G} \) which are roots of a polynomial in \( \mathbb{F}[x] \). Show \( \mathbb{A} \) is also countable.

11. This problem is on partial fractions. Suppose you have

\[ R(x) = \frac{p(x)}{q_1(x) \cdots q_m(x)}, \] degree of \( p(x) < \) degree of denominator.

where the polynomials \( q_i(x) \) are relatively prime and all the polynomials \( p(x) \) and \( q_i(x) \) have coefficients in a field of scalars \( \mathbb{F} \). Thus there exist polynomials \( a_i(x) \) having coefficients in \( \mathbb{F} \) such that

\[ 1 = \sum_{i=1}^{m} a_i(x) q_i(x) \]
3.5. EXERCISES

Explain why

\[ R(x) = \frac{p(x) \sum_{i=1}^{m} a_i(x) q_i(x)}{q_1(x) \cdots q_m(x)} = \sum_{i=1}^{m} \frac{a_i(x) p(x)}{\prod_{j \neq i} q_j(x)} \]

Now continue doing this on each term in the above sum till finally you obtain an expression of the form

\[ \sum_{i=1}^{m} \frac{b_i(x)}{q_i(x)} \]

Using the Euclidean algorithm for polynomials, explain why the above is of the form

\[ M(x) + \sum_{i=1}^{m} \frac{r_i(x)}{q_i(x)} \]

where the degree of each \( r_i(x) \) is less than the degree of \( q_i(x) \) and \( M(x) \) is a polynomial. Now argue that \( M(x) = 0 \). From this explain why the usual partial fractions expansion of calculus must be true. You can use the fact that every polynomial having real coefficients factors into a product of irreducible quadratic polynomials and linear polynomials having real coefficients. This follows from the fundamental theorem of algebra.

12. It was shown in the chapter that \( \mathbb{A} \) is a field. Here \( \mathbb{A} \) are the numbers in \( \mathbb{R} \) which are roots of a rational polynomial. Then it was shown in Problem 11 that it was actually countable. Show that \( \mathbb{A} + i\mathbb{A} \) is also an example of a countable field.
Chapter 4

Matrices

You have now solved systems of equations by writing them in terms of an augmented matrix and then doing row operations on this augmented matrix. It turns out that such rectangular arrays of numbers are important from many other different points of view. Numbers are also called scalars. In general, scalars are just elements of some field.

A matrix is a rectangular array of numbers from a field $\mathbb{F}$. For example, here is a matrix.

$$
\begin{pmatrix}
1 & 2 & 3 & 4 \\
5 & 2 & 8 & 7 \\
6 & -9 & 1 & 2 \\
\end{pmatrix}
$$

This matrix is a $3 \times 4$ matrix because there are three rows and four columns. The first row is $(1 \ 2 \ 3 \ 4)$, the second row is $(5 \ 2 \ 8 \ 7)$ and so forth. The first column is $\begin{pmatrix} 1 \\ 5 \\ 6 \end{pmatrix}$. The convention in dealing with matrices is to always list the rows first and then the columns. Also, you can remember the columns are like columns in a Greek temple. They stand up right while the rows just lie there like rows made by a tractor in a plowed field. Elements of the matrix are identified according to position in the matrix. For example, 8 is in position 2,3 because it is in the second row and the third column. You might remember that you always list the rows before the columns by using the phrase *Rowman Catholic*. The symbol, $(a_{ij})$ refers to a matrix in which the $i$ denotes the row and the $j$ denotes the column. Using this notation on the above matrix, $a_{23} = 8, a_{32} = -9, a_{12} = 2$, etc.

There are various operations which are done on matrices. They can sometimes be added, multiplied by a scalar and sometimes multiplied. To illustrate scalar multiplication, consider the following example.

$$
3 \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 2 & 8 & 7 \\ 6 & -9 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 6 & 9 & 12 \\ 15 & 6 & 24 & 21 \\ 18 & -27 & 3 & 6 \end{pmatrix}.
$$

The new matrix is obtained by multiplying every entry of the original matrix by the given scalar. If $A$ is an $m \times n$ matrix $-A$ is defined to equal $(-1)A$.

Two matrices which are the same size can be added. When this is done, the result is the matrix which is obtained by adding corresponding entries. Thus

$$
\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 2 \end{pmatrix} + \begin{pmatrix} -1 & 4 \\ 2 & 8 \\ 6 & -4 \end{pmatrix} = \begin{pmatrix} 0 & 6 \\ 5 & 12 \\ 11 & -2 \end{pmatrix}.
$$

Two matrices are equal exactly when they are the same size and the corresponding entries are
identical. Thus
\[
\begin{pmatrix}
0 & 0 \\
0 & 0 \\
0 & 0 \\
\end{pmatrix}
\neq
\begin{pmatrix}
0 & 0 \\
0 & 0 \\
\end{pmatrix}
\]
because they are different sizes. As noted above, you write \((c_{ij})\) for the matrix \(C\) whose \(ij^{th}\) entry is \(c_{ij}\). In doing arithmetic with matrices you must define what happens in terms of the \(c_{ij}\) sometimes called the entries of the matrix or the components of the matrix.

The above discussion stated for general matrices is given in the following definition.

**Definition 4.0.1** Let \(A = (a_{ij})\) and \(B = (b_{ij})\) be two \(m \times n\) matrices. Then \(A + B = C\) where
\[
C = (c_{ij})
\]
for \(c_{ij} = a_{ij} + b_{ij}\). Also if \(x\) is a scalar,
\[
x (a_{ij}) = (c_{ij})
\]
where \(c_{ij} = xa_{ij}\). The number \(A_{ij}\) will also typically refer to the \(ij^{th}\) entry of the matrix \(A\). The zero matrix, denoted by \(0\) will be the matrix consisting of all zeros.

Do not be upset by the use of the subscripts, \(ij\). The expression \(c_{ij} = a_{ij} + b_{ij}\) is just saying that you add corresponding entries to get the result of summing two matrices as discussed above.

Note that there are \(2 \times 3\) zero matrices, \(3 \times 4\) zero matrices, etc. In fact for every size there is a zero matrix.

With this definition, the following properties are all obvious but you should verify all of these properties are valid for \(A, B,\) and \(C; m \times n\) matrices and \(0\) an \(m \times n\) zero matrix,
\[
A + B = B + A,
\]
the commutative law of addition,
\[
(A + B) + C = A + (B + C),
\]
the associative law for addition,
\[
A + 0 = A,
\]
the existence of an additive identity,
\[
A + (−A) = 0,
\]
the existence of an additive inverse. Also, for \(α, β\) scalars, the following also hold.
\[
α (A + B) = αA + αB,
\]
\[
(α + β) A = αA + βA,
\]
\[
α (βA) = αβ (A),
\]
\[
1A = A.
\]

The above properties, **4.1 - 4.8** are the vector space axioms and the fact that the \(m \times n\) matrices satisfy these axioms is what is meant by saying this set of matrices with addition and scalar multiplication as defined above forms a vector space.

**Definition 4.0.2** Matrices which are \(n \times 1\) or \(1 \times n\) are especially called vectors and are often denoted by a bold letter. Thus
\[
\mathbf{x} = \begin{pmatrix} x_1 \\
\vdots \\
x_n \end{pmatrix}
\]
is an \(n \times 1\) matrix also called a column vector while a \(1 \times n\) matrix of the form \((x_1 \cdots x_n)\) is referred to as a row vector.
All the above is fine, but the real reason for considering matrices is that they can be multiplied. This is where things quit being banal. The following is the definition of multiplying a $m \times n$ matrix times a $n \times 1$ vector.

**Definition 4.0.3** Let $A = A_{ij}$ be an $m \times n$ matrix and let $v$ be an $n \times 1$ matrix,

$$v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}, \quad A = (a_1, \ldots, a_n)$$

where $a_i$ is an $m \times 1$ vector. Then $Av$, written as

$$\begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix},$$

is the $m \times 1$ column vector which equals the following linear combination of the columns.

$$v_1 a_1 + v_2 a_2 + \cdots + v_n a_n \equiv \sum_{j=1}^{n} v_j a_j \quad (4.9)$$

If the $j^{th}$ column of $A$ is

$$\begin{pmatrix} A_{1j} \\ A_{2j} \\ \vdots \\ A_{mj} \end{pmatrix}$$

then (4.9) takes the form

$$v_1 \begin{pmatrix} A_{11} \\ A_{21} \\ \vdots \\ A_{m1} \end{pmatrix} + v_2 \begin{pmatrix} A_{12} \\ A_{22} \\ \vdots \\ A_{m2} \end{pmatrix} + \cdots + v_n \begin{pmatrix} A_{1n} \\ A_{2n} \\ \vdots \\ A_{mn} \end{pmatrix}$$

Thus the $i^{th}$ entry of $Av$ is $\sum_{j=1}^{n} A_{ij} v_j$. Note that multiplication by an $m \times n$ matrix takes an $n \times 1$ matrix, and produces an $m \times 1$ matrix (vector).

Here is another example.

**Example 4.0.4** Compute the following product in $\mathbb{Z}_5$. That is, all the numbers are interpreted as residue classes.

$$\begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 2 & 1 & 3 \\ 2 & 1 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 4 \\ 1 \end{pmatrix}.$$

It equals

$$1 \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + 2 \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} + 4 \begin{pmatrix} 1 \\ 4 \\ 4 \end{pmatrix} + 1 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + \begin{pmatrix} 4 \\ 4 \\ 3 \end{pmatrix} + \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$
Example 4.0.5 Do the same problem when the numbers are just ordinary real numbers. That is, do it in $\mathbb{R}$.

You can work this out as follows.

\[
\begin{pmatrix}
1 & 2 & 1 \\
0 & 2 & 1 \\
2 & 0 & 1
\end{pmatrix} + 2 \begin{pmatrix}
1 & 2 & 1 \\
1 & 2 & 1 \\
1 & 2 & 1
\end{pmatrix} + 4 \begin{pmatrix}
1 & 3 \\
4 & 3 \\
1 & 1
\end{pmatrix} = 
\begin{pmatrix}
1 & 2 & 1 \\
0 & 2 & 1 \\
2 & 1 & 4
\end{pmatrix} + 4 \begin{pmatrix}
1 & 3 \\
4 & 3 \\
1 & 1
\end{pmatrix} = 
\begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
\]

With this done, the next task is to multiply an $m \times n$ matrix times an $n \times p$ matrix. Before doing so, the following may be helpful.

If the two middle numbers don’t match, you can’t multiply the matrices!

Definition 4.0.6 Let $A$ be an $m \times n$ matrix and let $B$ be an $n \times p$ matrix. Then $B$ is of the form

\[B = (b_1, \ldots, b_p)\]

where $b_k$ is an $n \times 1$ matrix. Then an $m \times p$ matrix $AB$ is defined as follows:

\[AB \equiv (Ab_1, \ldots, Ab_p)\]

(4.10)

where $Ab_k$ is an $m \times 1$ matrix. Hence $AB$ as just defined is an $m \times p$ matrix. For example,

Example 4.0.7 Multiply the following.

\[
\begin{pmatrix}
1 & 2 & 1 \\
0 & 2 & 1
\end{pmatrix} \begin{pmatrix}
1 & 2 & 0 \\
0 & 3 & 1 \\
-2 & 1 & 1
\end{pmatrix}
\]

The first thing you need to check before doing anything else is whether it is possible to do the multiplication. The first matrix is a $2 \times 3$ and the second matrix is a $3 \times 3$. Therefore, it is possible to multiply these matrices. According to the above discussion it should be a $2 \times 3$ matrix of the form

\[
\begin{pmatrix}
1 & 2 & 1 \\
0 & 2 & 1
\end{pmatrix} \begin{pmatrix}
1 & 2 & 1 \\
0 & 2 & 1 \\
-2 & 1 & 1
\end{pmatrix} = (Ab_1, \ldots, Ab_p)
\]

You know how to multiply a matrix times a vector and so you do so to obtain each of the three columns. Thus

\[
\begin{pmatrix}
1 & 2 & 1 \\
0 & 2 & 1
\end{pmatrix} \begin{pmatrix}
1 & 2 & 0 \\
0 & 3 & 1 \\
-2 & 1 & 1
\end{pmatrix} = \begin{pmatrix}
-1 & 9 & 3 \\
-2 & 7 & 3
\end{pmatrix}.
\]

Here is another example.
Example 4.0.8 Multiply the following.

\[
\begin{pmatrix}
1 & 2 & 0 \\
0 & 3 & 1 \\
-2 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 2 & 1 \\
0 & 2 & 1
\end{pmatrix}
\]

First check if it is possible. This is of the form \((3 \times 3)(2 \times 3)\). The inside numbers do not match and so you can’t do this multiplication. This means that anything you write will be absolute nonsense because it is impossible to multiply these matrices in this order. Aren’t they the same two matrices considered in the previous example? Yes they are. It is just that here they are in a different order. This shows something you must always remember about matrix multiplication.

Order Matters!

Matrix multiplication is not commutative. This is very different than multiplication of numbers!

4.1 The Entries of a Product

It is important to describe matrix multiplication in terms of entries of the matrices. What is the \(ij^{th}\) entry of \(AB\)? It would be the \(i^{th}\) entry of the \(j^{th}\) column of \(AB\). Thus it would be the \(i^{th}\) entry of \(Ab_j\). Now

\[
b_j = \begin{pmatrix} B_{1j} \\ \vdots \\ B_{nj} \end{pmatrix}
\]

and from the above definition, the \(i^{th}\) entry is

\[
\sum_{k=1}^{n} A_{ik} B_{kj}.
\]

In terms of pictures of the matrix, you are doing

\[
\begin{pmatrix}
A_{11} & A_{12} & \cdots & A_{1n} \\
A_{21} & A_{22} & \cdots & A_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
A_{m1} & A_{m2} & \cdots & A_{mn}
\end{pmatrix}
\begin{pmatrix}
B_{11} & B_{12} & \cdots & B_{1p} \\
B_{21} & B_{22} & \cdots & B_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
B_{n1} & B_{n2} & \cdots & B_{np}
\end{pmatrix}
\]

Then as explained above, the \(j^{th}\) column is of the form

\[
\begin{pmatrix}
A_{11} & A_{12} & \cdots & A_{1n} \\
A_{21} & A_{22} & \cdots & A_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
A_{m1} & A_{m2} & \cdots & A_{mn}
\end{pmatrix}
\begin{pmatrix} B_{1j} \\ B_{2j} \\ \vdots \\ B_{nj} \end{pmatrix}
\]

which is a \(m \times 1\) matrix or column vector which equals

\[
\begin{pmatrix}
A_{11} \\
A_{21} \\
\vdots \\
A_{m1}
\end{pmatrix} B_{ij} + \begin{pmatrix}
A_{12} \\
A_{22} \\
\vdots \\
A_{m2}
\end{pmatrix} B_{2j} + \cdots + \begin{pmatrix}
A_{1n} \\
A_{2n} \\
\vdots \\
A_{mn}
\end{pmatrix} B_{nj}.
\]
The $i^{th}$ entry of this $m \times 1$ matrix is

$$A_{i1}B_{1j} + A_{i2}B_{2j} + \cdots + A_{in}B_{nj} = \sum_{k=1}^{m} A_{ik}B_{kj}.$$  

This shows the following definition for matrix multiplication in terms of the $ij^{th}$ entries of the product harmonizes with Definition 4.0.3.

This motivates the definition for matrix multiplication which identifies the $ij^{th}$ entries of the product.

**Definition 4.1.1** Let $A = (A_{ij})$ be an $m \times n$ matrix and let $B = (B_{ij})$ be an $n \times p$ matrix. Then $AB$ is an $m \times p$ matrix and

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}. \quad (4.12)$$

Two matrices, $A$ and $B$ are said to be conformable in a particular order if they can be multiplied in that order. Thus if $A$ is an $r \times s$ matrix and $B$ is a $s \times p$ then $A$ and $B$ are conformable in the order $AB$. The above formula for $(AB)_{ij}$ says that it equals the $i^{th}$ row of $A$ times the $j^{th}$ column of $B$.

**Example 4.1.2** Multiply if possible

$$\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \text{ and } \begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \end{pmatrix} \text{ in } \mathbb{Z}_{11}.$$  

First check to see if this is possible. It is of the form $(3 \times 2)(2 \times 3)$ and since the inside numbers match, it must be possible to do this and the result should be a $3 \times 3$ matrix. The answer is of the form

$$\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \cdot \begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \end{pmatrix} = \begin{pmatrix} 5 & 4 & 5 \\ 2 & 4 & 5 \\ 2 & 9 & 3 \end{pmatrix},$$

where the commas separate the columns in the resulting product. Thus the above product equals a $3 \times 3$ matrix as desired.

**Example 4.1.3** Multiply if possible

$$\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \text{ and } \begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \end{pmatrix}.$$  

This is not possible because it is of the form $(3 \times 2)(3 \times 3)$ and the middle numbers don’t match.

**Example 4.1.4** Multiply if possible

$$\begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix}.$$  

This is possible because in this case it is of the form $(3 \times 3)(3 \times 2)$ and the middle numbers do match. When the multiplication is done it equals

$$\begin{pmatrix} 13 & 13 \\ 29 & 32 \\ 0 & 0 \end{pmatrix}.$$  

Check this and be sure you come up with the same answer.
Example 4.1.5 Multiply if possible \[
\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 & 0 \\ 1 & 2 & 1 & 0 \end{pmatrix}.
\]

In this case you are trying to do \((3 \times 1)(1 \times 4)\). The inside numbers match so you can do it. Verify
\[
\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 & 0 \\ 1 & 2 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 2 & 4 & 2 & 0 \\ 1 & 2 & 1 & 0 \end{pmatrix}
\]

4.2 Properties Of Matrix Multiplication

As pointed out above, sometimes it is possible to multiply matrices in one order but not in the other order. What if it makes sense to multiply them in either order? Will they be equal then?

Example 4.2.1 Compare \[
\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.
\]

The first product is
\[
\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 4 & 3 \end{pmatrix},
\]
the second product is
\[
\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix},
\]
and you see these are not equal. Therefore, you cannot conclude that \(AB = BA\) for matrix multiplication. However, there are some properties which do hold.

Proposition 4.2.2 If all multiplications and additions make sense, the following hold for matrices, \(A, B, C\) and \(a, b\) scalars.
\[
A (aB + bC) = a (AB) + b (AC) \quad (4.13)
\]
\[
(B + C) A = BA + CA \quad (4.14)
\]
\[
A (BC) = (AB) C \quad (4.15)
\]

Proof: Using the above definition of matrix multiplication,
\[
(A (aB + bC))_{ij} = \sum_k A_{ik} (aB + bC)_{kj}
\]
\[
= \sum_k A_{ik} (aB_{kj} + bC_{kj})
\]
\[
= a \sum_k A_{ik} B_{kj} + b \sum_k A_{ik} C_{kj}
\]
\[
= a (AB)_{ij} + b (AC)_{ij}
\]
\[
= (a (AB) + b (AC))_{ij}
\]
showing that \(A (B + C) = AB + AC\) as claimed. Formula (4.14) is entirely similar.
Consider the associative law of multiplication. Before reading this, review the definition of matrix multiplication in terms of entries of the matrices.

\[
(A(BC))_{ij} = \sum_k A_{ik} (BC)_{kj} = \sum_k A_{ik} \sum_l B_{kl} C_{lj} = \sum_l (AB)_{il} C_{lj} = ((AB)C)_{ij} \]

Another important operation on matrices is that of taking the transpose. The following example shows what is meant by this operation, denoted by placing a \( T \) as an exponent on the matrix.

\[
\begin{pmatrix}
1 & 1 + 2i \\
3 & 1 \\
2 & 6
\end{pmatrix}^T = \begin{pmatrix}
1 & 3 & 2 \\
1 + 2i & 1 & 6
\end{pmatrix}
\]

What happened? The first column became the first row and the second column became the second row. Thus the \( 3 \times 2 \) matrix became a \( 2 \times 3 \) matrix. The number 3 was in the second row and the first column and it ended up in the first row and second column. This motivates the following definition of the transpose of a matrix.

**Definition 4.2.3** Let \( A \) be an \( m \times n \) matrix. Then \( A^T \) denotes the \( n \times m \) matrix which is defined as follows.

\[(A^T)_{ij} = A_{ji}\]

The transpose of a matrix has the following important property.

**Lemma 4.2.4** Let \( A \) be an \( m \times n \) matrix and let \( B \) be a \( n \times p \) matrix. Then

\[(AB)^T = B^T A^T \quad (4.16)\]

and if \( \alpha \) and \( \beta \) are scalars,

\[(\alpha A + \beta B)^T = \alpha A^T + \beta B^T \quad (4.17)\]

**Proof:** From the definition,

\[
\left( (AB)^T \right)_{ij} = (AB)_{ji} = \sum_k A_{jk} B_{ki} = \sum_k (B^T)_{ik} (A^T)_{kj} = (B^T A^T)_{ij}
\]

is left as an exercise. \( \blacksquare \)

**Definition 4.2.5** An \( n \times n \) matrix \( A \) is said to be symmetric if \( A = A^T \). It is said to be skew symmetric if \( A^T = -A \).

**Example 4.2.6** Let

\[
A = \begin{pmatrix}
2 & 1 & 3 \\
1 & 5 & -3 \\
3 & -3 & 7
\end{pmatrix}
\]

Then \( A \) is symmetric.
Example 4.2.7 Let

\[ A = \begin{pmatrix} 0 & 1 & 3 \\ -1 & 0 & 2 \\ -3 & -2 & 0 \end{pmatrix} \]

Then \( A \) is skew symmetric.

There is a special matrix called \( I \) and defined by \( I_{ij} = \delta_{ij} \) where \( \delta_{ij} \) is the Kronecker symbol defined by

\[ \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \]

It is called the identity matrix because it is a multiplicative identity in the following sense.

Lemma 4.2.8 Suppose \( A \) is an \( m \times n \) matrix and \( I_n \) is the \( n \times n \) identity matrix. Then \( AI_n = A \). If \( I_m \) is the \( m \times m \) identity matrix, it also follows that \( I_mA = A \).

Proof:

\[(AI_n)_{ij} = \sum_k A_{ik} \delta_{kj} = A_{ij}\]

and so \( AI_n = A \). The other case is left as an exercise for you.

Definition 4.2.9 An \( n \times n \) matrix \( A \) has an inverse \( A^{-1} \) if and only if there exists a matrix, denoted as \( A^{-1} \) such that \( AA^{-1} = A^{-1}A = I \) where \( I = (\delta_{ij}) \) for

\[ \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \]

Such a matrix is called invertible.

If it acts like an inverse, then it is the inverse. This is the message of the following proposition.

Proposition 4.2.10 Suppose \( AB = BA = I \). Then \( B = A^{-1} \).

Proof: From the definition, \( B \) is an inverse for \( A \). Could there be another one \( B' \)?

\[ B' = B'I = B'(AB) = (B'A)B = IB = B. \]

Thus, the inverse, if it exists, is unique. ■

4.3 Finding The Inverse Of A Matrix

A little later a formula is given for the inverse of a matrix. However, it is not a good way to find the inverse for a matrix. There is a much easier way and it is this which is presented here. It is also important to note that not all matrices have inverses.

Example 4.3.1 Let \( A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \). Does \( A \) have an inverse?

One might think \( A \) would have an inverse because it does not equal zero. However,

\[ \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \]

and if \( A^{-1} \) existed, this could not happen because you could multiply on the left by the inverse \( A \) and conclude the vector \((-1, 1)^T = (0, 0)^T \). Thus the answer is that \( A \) does not have an inverse.
Suppose you want to find $B$ such that $AB = I$. Let

$$B = \begin{pmatrix} b_1 & \cdots & b_n \end{pmatrix}$$

Also the $i^{th}$ column of $I$ is

$$e_i = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}^T$$

Thus, if $AB = I$, $b_i$, the $i^{th}$ column of $B$ must satisfy the equation $A b_i = e_i$. The augmented matrix for finding $b_i$ is $(A | e_i)$. Thus, by doing row operations till $A$ becomes $I$, you end up with $(I | b_i)$ where $b_i$ is the solution to $A b_i = e_i$. Now the same sequence of row operations works regardless of the right side of the augmented matrix $(A | e_i)$ and so you can save trouble by simply doing the following.

$$(A | I) \xrightarrow{\text{row operations}} (I | B)$$

and the $i^{th}$ column of $B$ is $b_i$, the solution to $A b_i = e_i$. Thus $AB = I$.

This is the reason for the following simple procedure for finding the inverse of a matrix. This procedure is called the Gauss Jordan procedure. It produces the inverse if the matrix has one. Actually, it produces the right inverse.

**Procedure 4.3.2** Suppose $A$ is an $n \times n$ matrix. To find $A^{-1}$ if it exists, form the augmented $n \times 2n$ matrix,

$$(A | I)$$

and then do row operations until you obtain an $n \times 2n$ matrix of the form

$$(I | B)$$

if possible. When this has been done, $B = A^{-1}$. The matrix $A$ has an inverse exactly when it is possible to do row operations and end up with one like $I$.

Here is a fundamental theorem which describes when a matrix has an inverse.

**Theorem 4.3.3** Let $A$ be an $n \times n$ matrix. Then $A^{-1}$ exists if and only if the columns of $A$ are a linearly independent set. Also, if $A$ has a right inverse, then it has an inverse which equals the right inverse.

**Proof:** $\Rightarrow$ If $A^{-1}$ exists, then $A^{-1} A = I$ and so $A x = 0$ if and only if $x = 0$. Why? But this says that the columns of $A$ are linearly independent.

$\Leftarrow$ Say the columns are linearly independent. Then there exists $b_i \in \mathbb{F}^n$ such that

$$A b_i = e_i$$

where $e_i$ is the column vector with 1 in the $i^{th}$ position and zeros elsewhere. Then from the way we multiply matrices,

$$A \begin{pmatrix} b_1 & \cdots & b_n \end{pmatrix} = \begin{pmatrix} e_1 & \cdots & e_n \end{pmatrix} = I$$

Thus $A$ has a right inverse. Now letting $B \equiv \begin{pmatrix} b_1 & \cdots & b_n \end{pmatrix}$, it follows that $B x = 0$ if and only if $x = 0$. However, this is nothing but a statement that the columns of $B$ are linearly independent. Hence, by what was just shown, $B$ has a right inverse $C$, $B C = I$. Then from $A B = I$, it follows that

$$A = A (B C) = (A B) C = I C = C$$

and so $AB = BC = BA = I$. Thus the inverse exists.

Finally, if $AB = I$, then $B x = 0$ if and only if $x = 0$ and so the columns of $B$ are a linearly independent set in $\mathbb{F}^n$. Therefore, it has a right inverse $C$ which by a repeat of the above argument is $A$. Thus $AB = BA = I$. $\blacksquare$

Similarly, if $A$ has a left inverse then it has an inverse which is the same as the left inverse.

The theorem gives a condition for the existence of the inverse and the above procedure gives a method for finding it.
Example 4.3.4 Let \( A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \). Find \( A^{-1} \) in arithmetic of \( \mathbb{Z}_3 \).

Form the augmented matrix
\[
\begin{pmatrix}
1 & 0 & 1 & 1 & 0 & 0 \\
1 & 2 & 1 & 0 & 1 & 0 \\
1 & 1 & 2 & 0 & 0 & 1
\end{pmatrix}.
\]

Now do row operations in \( \mathbb{Z}_3 \) until the \( n \times n \) matrix on the left becomes the identity matrix. This yields after some computations,
\[
\begin{pmatrix}
1 & 0 & 0 & 0 & 2 & 2 \\
0 & 1 & 0 & 1 & 2 & 0 \\
0 & 0 & 1 & 1 & 1 & 1
\end{pmatrix}
\]
and so the inverse of \( A \) is the matrix on the right,
\[
\begin{pmatrix}
0 & 2 & 2 \\
1 & 2 & 0 \\
1 & 1 & 1
\end{pmatrix}.
\]

Checking the answer is easy. Just multiply the matrices and see if it works.
\[
\begin{pmatrix}
1 & 0 & 1 \\
1 & 2 & 1 \\
1 & 1 & 2
\end{pmatrix}\begin{pmatrix}
0 & 2 & 2 \\
1 & 2 & 0 \\
1 & 1 & 1
\end{pmatrix} = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]
All arithmetic is done in \( \mathbb{Z}_3 \). Always check your answer because if you are like some of us, you will usually have made a mistake.

Example 4.3.5 Let \( A = \begin{pmatrix} 6 & -1 & 2 \\ -1 & 2 & -1 \\ 2 & -1 & 1 \end{pmatrix} \). Find \( A^{-1} \) in \( \mathbb{Q} \).

Set up the augmented matrix \((A|I)\)
\[
\begin{pmatrix}
6 & -1 & 2 & 1 & 0 & 0 \\
-1 & 2 & -1 & 0 & 1 & 0 \\
2 & -1 & 1 & 0 & 0 & 1
\end{pmatrix}
\]

Now find row reduced echelon form
\[
\begin{pmatrix}
1 & 0 & 0 & 1 & -1 & -3 \\
0 & 1 & 0 & -1 & 2 & 4 \\
0 & 0 & 1 & -3 & 4 & 11
\end{pmatrix}
\]
Thus the inverse is
\[
\begin{pmatrix}
1 & -1 & -3 \\
-1 & 2 & 4 \\
-3 & 4 & 11
\end{pmatrix}
\]

Example 4.3.6 Let \( A = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 2 & 2 & 4 \end{pmatrix} \). Find \( A^{-1} \).
This time there is no inverse because the columns are not linearly independent. This can be seen by solving the equation
\[
\begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 2 & 2 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
\]
and finding that there is a nonzero solution which is equivalent to the columns being a dependent set. Thus, by Theorem 4.3.3, there is no inverse.

**Example 4.3.7** Consider the matrix
\[
\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 5 \end{pmatrix}
\]
Find its inverse in arithmetic of \( \mathbb{Q} \) and then find its inverse in \( \mathbb{Z}_5 \).

It has an inverse in \( \mathbb{Q} \).
\[
\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 5 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{5} \end{pmatrix}
\]
However, in \( \mathbb{Z}_5 \) it has no inverse because 5 = 0 in \( \mathbb{Z}_5 \) and so in \( \mathbb{Z}_5^3 \), the columns are dependent.

**Example 4.3.8** Here is a matrix.
\[
\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}
\]
Find its inverse in the arithmetic of \( \mathbb{Q} \) and then in \( \mathbb{Z}_3 \).

It has an inverse in the arithmetic of \( \mathbb{Q} \)
\[
\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{pmatrix}
\]
However, there is no inverse in the arithmetic of \( \mathbb{Z}_3 \). Indeed, the row reduced echelon form of
\[
\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \end{pmatrix}
\]
computed in \( \mathbb{Z}_3 \) is
\[
\begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}
\]
and so \( \begin{pmatrix} 1 \\ 1 \end{pmatrix} \in \ker \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \) which shows that the columns are not independent so there is no inverse in \( \mathbb{Z}_3^2 \).

The field of residue classes is not of major importance in this book, but it is included to emphasize that these considerations are completely algebraic in nature, depending only on field axioms. There is no geometry or analysis involved here.
4.4 Matrices And Systems Of Equations

Suppose you have the following system of equations.

\[
\begin{align*}
x - 5w - 3z &= 1 \\
2w + x + y + z &= 2 \\
2w + x + y + z &= 3
\end{align*}
\]

You could write it in terms of matrix multiplication as follows.

\[
\begin{pmatrix}
1 & 0 & -3 & -5 \\
1 & 1 & 1 & 2 \\
1 & 1 & 1 & 2
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z \\
w
\end{pmatrix}
= 
\begin{pmatrix}
1 \\
2 \\
3
\end{pmatrix}
\]

You could also write it in terms of vector addition as follows.

\[
x \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + z \begin{pmatrix} -3 \\ 1 \\ 1 \end{pmatrix} + w \begin{pmatrix} -5 \\ 2 \\ 2 \end{pmatrix} = 
\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}
\]

When you find a solution to the system of equations, you are really finding the scalars so that the vector on the right is the above linear combination. In other words, you are finding a linear relationship between the last column and those before it.

We considered writing this as an augmented matrix

\[
\begin{pmatrix}
1 & 0 & -3 & -5 & 1 \\
1 & 1 & 1 & 2 & 2 \\
1 & 1 & 1 & 2 & 3
\end{pmatrix}
\]

and then row reducing it to get a matrix in row reduced echelon form from which it was easy to see the solution, finding the last column as a linear combination of the preceding columns. However, this process of row reduction works just as well to find the fourth column as a linear combination of the first three and the third as a linear combination of the first two, so when you reduce to row reduced echelon form, you are really solving many systems of equations at the same time. The important thing was the observation that the row operations did not change the solution set of the system. However, this could be said differently. The row operations did not change the available scalars in forming the last column as a linear combination of the first four. Similarly, the row operations did not change the available scalars to obtain the fourth column as a linear combination of the first three, and so forth. In other words, if a column is a linear combination of the preceding columns, then after doing row operations, that column will still be the same linear combination of the preceding columns. Thus we have the following significant observation which is stated here as a theorem.

**Theorem 4.4.1** Row operations preserve all linear relationships between columns.

Now here is a slightly different description of the row reduced echelon form.

**Definition 4.4.2** Let \( e_i \) denote the column vector which has all zero entries except for the \( i^{th} \) slot which is one. An \( m \times n \) matrix is said to be in row reduced echelon form if, in viewing successive columns from left to right, the first nonzero column encountered is \( e_1 \) and if you have encountered \( e_1, e_2, \ldots, e_k \), the next column is either \( e_{k+1} \) or is a linear combination of the vectors, \( e_1, e_2, \ldots, e_k \).

Earlier an algorithm was presented which will produce a matrix in row reduced echelon form. A natural question is whether there is only one row reduced echelon form. In fact, there is only one and this follows easily from the above definition.
Suppose you had two $B, C$ in row reduced echelon form and these came from the same matrix $A$ through row operations. Then they have zero columns in the same positions because row operations preserve all zero columns. Also $B, C$ have $e_1$ in the same position because its position is that of the first column of $A$ which is not zero. Similarly $e_2, e_3$ and so forth must be in the same positions because of the above definition where these positions are defined in terms of a column being the first in $A$ when viewed from the left to the right which is not a linear combination of the columns before it. As to a column after $e_k$ and before $e_{k+1}$ if there is such, these are determined by the scalars which give this column in $A$ as a linear combination of the columns to its left because all linear relationships between columns are preserved by doing row operations. Thus $B, C$ must be exactly the same. This is why there is only one row reduced echelon form for a given matrix and justifies the use of the definite article when referring to it.

This proves the following theorem.

**Theorem 4.4.3** The row reduced echelon form is unique.

Now from this theorem, we can obtain the following.

**Theorem 4.4.4** Let $A$ be an $n \times n$ matrix. Then it is invertible if and only if there is a sequence of row operations which produces $I$.

**Proof:** $\Rightarrow$ Since $A$ is invertible, it follows from Theorem 4.3.3 that the columns of $A$ must be independent. Hence, in the row reduced echelon form for $A$, the columns must be $e_1, e_2, \ldots, e_n$ in order from left to right. In other words, there is a sequence of row operations which produces $I$.

$\Leftarrow$ Now suppose such a sequence of row operations produces $I$. Then since row operations preserve linear combinations between columns, it follows that no column is a linear combination of the others and consequently the columns are linearly independent. By Theorem 4.3.3 again, $A$ is invertible.

It would be possible to define things like rank in terms of the row reduced echelon form and this is often done. However, in this book, these things will be defined in terms of vector space language.

**Definition 4.4.5** Let $A$ be an $m \times n$ matrix, the entries being in $\mathbb{F}$ a field. Then rank $(A)$ is defined as the dimension of $\text{Im} (A) \equiv A (\mathbb{F}^n)$. Note that, from the way we multiply matrices times a vector, this is just the same as the dimension of span (columns of $A$), sometimes called the column space.

Now here is a very useful result.

**Proposition 4.4.6** Let $A$ be an $m \times n$ matrix. Then rank $(A)$ equals the number of pivot columns in the row reduced echelon form of $A$.

**Proof:** This is obvious if the matrix is already in row reduced echelon form. In this case, the pivot columns consist of $e_1, e_2, \ldots, e_r$ and every other column is a linear combination of these. Thus the rank of this matrix is $r$ because these vectors are obviously linearly independent. However, the linear relationships between a column and its preceding columns are preserved by row operations and so the columns in $A$ corresponding to the first occurrence of $e_1$, first occurrence of $e_2$ and so forth, in the row reduced echelon form, the pivot columns, are also a basis for the span of the columns of $A$ and so there are $r$ of these.

Note that from the description of the row reduced echelon form, the rank is also equal to the number of nonzero rows in the row reduced echelon form.

### 4.5 Block Multiplication Of Matrices

Consider the following problem

\[
\begin{pmatrix}
A & B \\
C & D
\end{pmatrix}
\begin{pmatrix}
E & F \\
G & H
\end{pmatrix}
\]
You know how to do this. You get
\[
\begin{pmatrix}
AE + BG & AF + BH \\
CE + DG & CF + DH
\end{pmatrix}.
\]

Now what if instead of numbers, the entries, A, B, C, D, E, F, G are matrices of a size such that the multiplications and additions needed in the above formula all make sense. Would the formula be true in this case? I will show below that this is true.

Suppose A is a matrix of the form
\[
A = \begin{pmatrix}
A_{11} & \cdots & A_{1m} \\
\vdots & \ddots & \vdots \\
A_{r1} & \cdots & A_{rm}
\end{pmatrix}
\]
(4.19)
where \(A_{ij}\) is a \(s_i \times p_j\) matrix where \(s_i\) is constant for \(j = 1, \cdots, m\) for each \(i = 1, \cdots, r\). Such a matrix is called a block matrix, also a partitioned matrix. How do you get the block \(A_{ij}\)? Here is how for \(A\) an \(m \times n\) matrix:

\[
\begin{pmatrix}
0 & I_{s_i \times s_i} & 0 \\
0 & I_{s_i \times s_i} & 0
\end{pmatrix}A
\begin{pmatrix}
0 & I_{p_j \times p_j} \\
0 & I_{p_j \times p_j}
\end{pmatrix}.
\]
(4.20)

In the block column matrix on the right, you need to have \(c_j - 1\) rows of zeros above the small \(p_j \times p_j\) identity matrix where the columns of \(A\) involved in \(A_{ij}\) are \(c_j, \cdots, c_j + p_j - 1\) and in the block row matrix on the left, you need to have \(r_i - 1\) columns of zeros to the left of the \(s_i \times s_i\) identity matrix where the rows of \(A\) involved in \(A_{ij}\) are \(r_i, \cdots, r_i + s_i\). An important observation to make is that the matrix on the right specifies columns to use in the block and the one on the left specifies the rows used. Thus the block \(A_{ij}\) in this case is a matrix of size \(s_i \times p_j\). There is no overlap between the blocks of \(A\). Thus the identity \(n \times n\) identity matrix corresponding to multiplication on the right of \(A\) is of the form
\[
\begin{pmatrix}
I_{p_1 \times p_1} & 0 & \cdots \\
0 & \ddots & \vdots \\
0 & \cdots & I_{p_m \times p_m}
\end{pmatrix}
\]
these little identity matrices don’t overlap. A similar conclusion follows from consideration of the matrices \(I_{s_i \times s_i}\).

Next consider the question of multiplication of two block matrices. Let \(B\) be a block matrix of the form
\[
\begin{pmatrix}
B_{11} & \cdots & B_{1p} \\
\vdots & \ddots & \vdots \\
B_{r1} & \cdots & B_{rp}
\end{pmatrix}
\]
(4.21)
and \(A\) is a block matrix of the form
\[
\begin{pmatrix}
A_{11} & \cdots & A_{1m} \\
\vdots & \ddots & \vdots \\
A_{p1} & \cdots & A_{pm}
\end{pmatrix}
\]
(4.22)
and that for all \(i, j\), it makes sense to multiply \(B_{is}A_{sj}\) for all \(s \in \{1, \cdots, p\}\). (That is the two matrices, \(B_{is}\) and \(A_{sj}\) are conformable,) and that for fixed \(ij\), it follows \(B_{is}A_{sj}\) is the same size for each \(s\) so that it makes sense to write \(\sum_s B_{is}A_{sj}\).
The following theorem says essentially that when you take the product of two matrices, you can do it two ways. One way is to simply multiply them forming $BA$. The other way is to partition both matrices, formally multiply the blocks to get another block matrix and this one will be $BA$ partitioned. Before presenting this theorem, here is a simple lemma which is really a special case of the theorem.

**Lemma 4.5.1** Consider the following product.

$$
\begin{pmatrix}
0 \\
I \\
0
\end{pmatrix}
\begin{pmatrix}
0 & I & 0 \\
0 & 0 & 0 \\
0 & I & 0
\end{pmatrix}
$$

where the first is $n \times r$ and the second is $r \times n$. The small identity matrix $I$ is an $r \times r$ matrix and there are $l$ zero rows above $I$ and $l$ zero columns to the left of $I$ in the right matrix. Then the product of these matrices is a block matrix of the form

$$
\begin{pmatrix}
0 & 0 & 0 \\
0 & I & 0 \\
0 & 0 & 0
\end{pmatrix}
$$

**Proof:** From the definition of the way you multiply matrices, the product is

$$
\begin{pmatrix}
0 & I & 0 \\
0 & 0 & 0 \\
0 & I & 0
\end{pmatrix}
\begin{pmatrix}
0 & 0 & 0 \\
0 & I & 0 \\
0 & 0 & 0
\end{pmatrix} \cdot
\begin{pmatrix}
e_1 & \cdots & e_r
\end{pmatrix}
\begin{pmatrix}
0 & 0 & 0 \\
0 & I & 0 \\
0 & 0 & 0
\end{pmatrix}
$$

which yields the claimed result. In the formula $e_j$ refers to the column vector of length $r$ which has a 1 in the $j$th position. ■

**Theorem 4.5.2** Let $B$ be a $q \times p$ block matrix as in 4.21 and let $A$ be a $p \times n$ block matrix as in 4.22 such that $B_{is}$ is conformable with $A_{sj}$ and each product, $B_{is}A_{sj}$ for $s = 1, \cdots, p$ is of the same size so they can be added. Then $BA$ can be obtained as a block matrix such that the $ij$th block is of the form

$$
\sum_s B_{is}A_{sj}.
$$

(4.23)

**Proof:** From 4.20

$$
B_{is}A_{sj} = \begin{pmatrix}
0 & I_{r_s \times r_i} & 0
\end{pmatrix} B \begin{pmatrix}
0 & I_{p_s \times p_s} \\
0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
0 & I_{p_s \times p_s} & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
A \begin{pmatrix}
0 & I_{q_j \times q_j}
\end{pmatrix}
$$

where here it is assumed $B_{is}$ is $r_s \times p_s$ and $A_{sj}$ is $p_s \times q_j$. The product involves the $s$th block in the $i$th row of blocks for $B$ and the $s$th block in the $j$th column of $A$. Thus there are the same number of rows above the $I_{p_s \times p_s}$ as there are columns to the left of $I_{p_s \times p_s}$ in those two inside matrices. Then from Lemma 4.5.1

$$
\begin{pmatrix}
0 \\
I_{p_s \times p_s} \\
0
\end{pmatrix}
\begin{pmatrix}
0 & I_{p_s \times p_s} & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
= I_{p_1 \times p_1}
\begin{pmatrix}
0 & \cdots & 0 \\
0 & \cdots & 0 \\
0 & \cdots & 0
\end{pmatrix}
$$

Since the blocks of small identity matrices do not overlap,

$$
\sum_s \begin{pmatrix}
0 & 0 & 0 \\
0 & I_{p_s \times p_s} & 0 \\
0 & 0 & 0
\end{pmatrix} = I_{p \times p}
$$
and so $\sum_s B_is As_j =$

$$\sum_s \begin{pmatrix} 0 & I_{r_i \times r_i} & 0 \\ I_{p_s \times p_s} & 0 \\ 0 & 0 & 0 \end{pmatrix} B \begin{pmatrix} 0 & I_{p_s \times p_s} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} A \begin{pmatrix} 0 & I_{q_j \times q_j} \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & I_{r_i \times r_i} & 0 \\ 0 & 0 & 0 \end{pmatrix} B \sum_s \begin{pmatrix} 0 & I_{p_s \times p_s} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} A \begin{pmatrix} 0 & I_{q_j \times q_j} \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & I_{r_i \times r_i} & 0 \\ 0 & 0 & 0 \end{pmatrix} BIA \begin{pmatrix} 0 & I_{q_j \times q_j} \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & I_{r_i \times r_i} & 0 \\ 0 & 0 & 0 \end{pmatrix} BA \begin{pmatrix} 0 & I_{q_j \times q_j} \\ 0 & 0 \end{pmatrix}$$

which equals the $ij^{th}$ block of $BA$. Hence the $ij^{th}$ block of $BA$ equals the formal multiplication according to matrix multiplication, $\sum_s B_is As_j$. ■

**Example 4.5.3** Let an $n \times n$ matrix have the form

$$A = \begin{pmatrix} a & b \\ c & P \end{pmatrix}$$

where $P$ is $n - 1 \times n - 1$. Multiply it by

$$B = \begin{pmatrix} p & q \\ r & Q \end{pmatrix}$$

where $B$ is also an $n \times n$ matrix and $Q$ is $n - 1 \times n - 1$.

You use block multiplication

$$\begin{pmatrix} a & b \\ c & P \end{pmatrix} \begin{pmatrix} p & q \\ r & Q \end{pmatrix} = \begin{pmatrix} ap + br & aq + bQ \\ pc + Pr & cq + PQ \end{pmatrix}$$

Note that this all makes sense. For example, $b = 1 \times n - 1$ and $r = n - 1 \times 1$ so $br$ is a $1 \times 1$. Similar considerations apply to the other blocks.

Here is a very significant application. A matrix is called block diagonal if it has all zeros except for square blocks down the diagonal. That is, it is of the form

$$A = \begin{pmatrix} A_1 & 0 \\ & A_2 \\ & & \ddots \\ 0 & & & A_m \end{pmatrix}$$

where $A_j$ is a $r_j \times r_j$ matrix whose main diagonal lies on the main diagonal of $A$. Then by block multiplication, if $p \in \mathbb{N}$ the positive integers,

$$A^p = \begin{pmatrix} A_1^p & 0 \\ & A_2^p \\ & & \ddots \\ 0 & & & A_m^p \end{pmatrix} \quad (4.24)$$

Also, $A^{-1}$ exists if and only if each block is invertible and in fact, $A^{-1}$ is given by the above when $p = -1$. 
4.6 Elementary Matrices

The elementary matrices result from doing a row operation to the identity matrix. Recall the following definition.

**Definition 4.6.1** The row operations consist of the following

1. Switch two rows.
2. Multiply a row by a nonzero number.
3. Replace a row by a multiple of another row added to it.

The elementary matrices are given in the following definition.

**Definition 4.6.2** The elementary matrices consist of those matrices which result by applying a single row operation to an identity matrix. Those which involve switching rows of the identity are called permutation matrices.\(^1\)

As an example of why these elementary matrices are interesting, consider the following.

\[
\begin{pmatrix}
0 & 1 & 0 \\
1 & 0 & 0 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
a & b & c & d \\
x & y & z & w \\
f & g & h & i
\end{pmatrix}
= 
\begin{pmatrix}
x & y & z & w \\
a & b & c & d \\
f & g & h & i
\end{pmatrix}
\]

A 3 \times 4 matrix was multiplied on the left by an elementary matrix which was obtained from row operation 1 applied to the identity matrix. This resulted in applying the operation 1 to the given matrix. This is what happens in general.

Now consider what these elementary matrices look like. First consider the one which involves switching row \(i\) and row \(j\) where \(i < j\). This matrix is of the form

\[
\begin{pmatrix}
1 & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & 1 \\
1 & 0 & \cdots & \ddots \\
0 & \cdots & \ddots & 1
\end{pmatrix}
\]

Note how the \(i^{th}\) and \(j^{th}\) rows are switched in the identity matrix and there are thus all ones on the main diagonal except for those two positions indicated. The two exceptional rows are shown. The \(i^{th}\) row was the \(j^{th}\) and the \(j^{th}\) row was the \(i^{th}\) in the identity matrix. Now consider what this does to a column vector.

\[
\begin{pmatrix}
x_1 \\
\vdots \\
x_i \\
\vdots \\
x_j \\
\vdots \\
x_n
\end{pmatrix}
\begin{pmatrix}
1 & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & 1 \\
1 & 0 & \cdots & \ddots \\
0 & \cdots & \ddots & 1
\end{pmatrix}
= 
\begin{pmatrix}
x_1 \\
\vdots \\
x_i \\
\vdots \\
x_j \\
\vdots \\
x_n
\end{pmatrix}
\]

\(^1\)More generally, a permutation matrix is a matrix which comes by permuting the rows of the identity matrix, not just switching two rows.
Now denote by $P^{ij}$ the elementary matrix which comes from the identity from switching rows $i$ and $j$. From what was just explained consider multiplication on the left by this elementary matrix.

$$
\begin{pmatrix}
 a_{11} & a_{12} & \cdots & a_{1p} \\
 \vdots & \vdots & & \vdots \\
 a_{i1} & a_{i2} & \cdots & a_{ip} \\
 \vdots & \vdots & & \vdots \\
 a_{j1} & a_{j2} & \cdots & a_{jp} \\
 \vdots & \vdots & & \vdots \\
 a_{n1} & a_{n2} & \cdots & a_{np}
\end{pmatrix}
$$

From the way you multiply matrices this is a matrix which has the indicated columns.

$$
\begin{pmatrix}
 a_{11} \\
 \vdots \\
 a_{i1} \\
 \vdots \\
 a_{j1} \\
 \vdots \\
 a_{n1}
\end{pmatrix}
, \begin{pmatrix}
 a_{12} \\
 \vdots \\
 a_{i2} \\
 \vdots \\
 a_{j2} \\
 \vdots \\
 a_{n2}
\end{pmatrix}
, \ldots, \begin{pmatrix}
 a_{1p} \\
 \vdots \\
 a_{ip} \\
 \vdots \\
 a_{jp} \\
 \vdots \\
 a_{np}
\end{pmatrix}
= \begin{pmatrix}
 a_{11} & a_{12} & \cdots & a_{1p} \\
 \vdots & \vdots & & \vdots \\
 a_{j1} & a_{j2} & \cdots & a_{jp} \\
 \vdots & \vdots & & \vdots \\
 a_{i1} & a_{i2} & \cdots & a_{ip} \\
 \vdots & \vdots & & \vdots \\
 a_{n1} & a_{n2} & \cdots & a_{np}
\end{pmatrix}
$$

This has established the following lemma.

**Lemma 4.6.3** Let $P^{ij}$ denote the elementary matrix which involves switching the $i^{th}$ and the $j^{th}$ rows. Then

$$
P^{ij} A = B
$$

where $B$ is obtained from $A$ by switching the $i^{th}$ and the $j^{th}$ rows.

Next consider the row operation which involves multiplying the $i^{th}$ row by a nonzero constant, $c$. The elementary matrix which results from applying this operation to the $i^{th}$ row of the identity matrix is of the form

$$
\begin{pmatrix}
 1 & 0 \\
 \ddots & \ddots \\
 1 & c \\
 \vdots & \ddots \\
 0 & \cdots & 1
\end{pmatrix}
$$
Now consider what this does to a column vector.

\[
\begin{pmatrix}
1 & 0 \\
\vdots & \\
1 & c \\
0 & 1
\end{pmatrix}
\begin{pmatrix}
v_1 \\
v_{i-1} \\
v_i \\
v_{i+1} \\
\vdots \\
v_n
\end{pmatrix} =
\begin{pmatrix}
v_1 \\
v_{i-1} \\
v_i \\
v_{i+1} \\
\vdots \\
v_n
\end{pmatrix}
\]

Denote by \( E(c,i) \) this elementary matrix which multiplies the \( i \)th row of the identity by the nonzero constant, \( c \). Then from what was just discussed and the way matrices are multiplied,

\[
E(c,i)
\begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1p} \\
a_{i1} & a_{i2} & \cdots & a_{ip} \\
a_{j1} & a_{j2} & \cdots & a_{jp} \\
a_{n1} & a_{n2} & \cdots & a_{np}
\end{pmatrix}
\]

equals a matrix having the columns indicated below.

\[
\begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1p} \\
a_{i1} & a_{i2} & \cdots & a_{ip} \\
a_{j1} & a_{j2} & \cdots & a_{jp} \\
a_{n1} & a_{n2} & \cdots & a_{np}
\end{pmatrix}
\begin{pmatrix}
a_{12} \\
a_{i2} \\
a_{j2} \\
a_{n2}
\end{pmatrix}
\cdots
\begin{pmatrix}
a_{1p} \\
a_{ip} \\
a_{jp} \\
a_{np}
\end{pmatrix}
\]

This proves the following lemma.

**Lemma 4.6.4** Let \( E(c,i) \) denote the elementary matrix corresponding to the row operation in which the \( i \)th row is multiplied by the nonzero constant \( c \). Thus \( E(c,i) \) involves multiplying the \( i \)th row of the identity matrix by \( c \). Then

\[
E(c,i)A = B
\]

where \( B \) is obtained from \( A \) by multiplying the \( i \)th row of \( A \) by \( c \).
Example 4.6.5  Consider this.
\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 5 & 0 \\
0 & 0 & 1 \\
\end{pmatrix}
\begin{pmatrix}
a & b \\
c & d \\
e & f \\
\end{pmatrix}
= 
\begin{pmatrix}
a & b \\
5c & 5d \\
e & f \\
\end{pmatrix}
\]

Finally consider the third of these row operations. Denote by \(E(c \times i + j)\) the elementary matrix which replaces the \(j^{th}\) row with itself added to \(c\) times the \(i^{th}\) row added to it. In case \(i < j\) this will be of the form
\[
\begin{pmatrix}
1 & & & \\
& \ddots & & \\
& & 1 & \\
& & & c \\
0 & & & 0 \\
\end{pmatrix}
\]

Now consider what this does to a column vector.
\[
\begin{pmatrix}
1 & & & \\
& \ddots & & \\
& & 1 & \\
& & & c \\
0 & & & 0 \\
\end{pmatrix}
\begin{pmatrix}
v_1 \\
v_i \\
v_j \\
v_n \\
\end{pmatrix}
= 
\begin{pmatrix}
v_1 \\
v_i \\
v_j \\
v_n \\
\end{pmatrix}
\]
The case where $i > j$ is handled similarly. This proves the following lemma.

**Lemma 4.6.6** Let $E(c \times i + j)$ denote the elementary matrix obtained from $I$ by replacing the $j^{th}$ row with $c$ times the $i^{th}$ row added to it. Then

$$E(c \times i + j) A = B$$

where $B$ is obtained from $A$ by replacing the $j^{th}$ row of $A$ with itself added to $c$ times the $i^{th}$ row of $A$.

**Example 4.6.7** Consider the third row operation.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \\ 2a + c & 2b + f \end{pmatrix}$$

The next theorem is the main result.

**Theorem 4.6.8** To perform any of the three row operations on a matrix $A$ it suffices to do the row operation on the identity matrix obtaining an elementary matrix $E$ and then take the product, $EA$.

Furthermore, each elementary matrix is invertible and its inverse is an elementary matrix.

**Proof:** The first part of this theorem has been proved in Lemmas 4.6.3 - 4.6.6. It only remains to verify the claim about the inverses. Consider first the elementary matrices corresponding to row operation of type three.

$$E(-c \times i + j) E(c \times i + j) = I$$

This follows because the first matrix takes $c$ times row $i$ in the identity and adds it to row $j$. When multiplied on the left by $E(-c \times i + j)$ it follows from the first part of this theorem that you take the $i^{th}$ row of $E(c \times i + j)$ which coincides with the $i^{th}$ row of $I$ since that row was not changed, multiply it by $-c$ and add to the $j^{th}$ row of $E(c \times i + j)$ which was the $j^{th}$ row of $I$ added to $c$ times the $i^{th}$ row of $I$. Thus $E(-c \times i + j)$ multiplied on the left, undoes the row operation which resulted in $E(c \times i + j)$. The same argument applied to the product

$$E(c \times i + j) E(-c \times i + j)$$

replacing $c$ with $-c$ in the argument yields that this product is also equal to $I$. Therefore,

$$E(c \times i + j)^{-1} = E(-c \times i + j).$$

Similar reasoning shows that for $E(c, i)$ the elementary matrix which comes from multiplying the $i^{th}$ row by the nonzero constant, $c$,

$$E(c, i)^{-1} = E(c^{-1}, i).$$

Finally, consider $P^{ij}$ which involves switching the $i^{th}$ and the $j^{th}$ rows.

$$P^{ij} P^{ij} = I$$
because by the first part of this theorem, multiplying on the left by $P_{ij}$ switches the $i^{th}$ and $j^{th}$ rows of $P_{ij}$ which was obtained from switching the $i^{th}$ and $j^{th}$ rows of the identity. First you switch them to get $P_{ij}$ and then you multiply on the left by $P_{ij}$ which switches these rows again and restores the identity matrix. Thus $(P_{ij})^{-1} = P_{ij}$. ■

Using Theorem 4.6.9, this shows the following result.

**Theorem 4.6.9** Let $A$ be an $n \times n$ matrix. Then if $R$ is its row reduced echelon form, there is a sequence of elementary matrices $E_i$ such that

$$E_1E_2\cdots E_mA = R$$

In particular, $A$ is invertible if and only if there is a sequence of elementary matrices as above such that

$$E_1E_2\cdots E_mA = I$$

Inverting these,

$$A = E_m^{-1}\cdots E_2^{-1}E_1^{-1}$$

a product of elementary matrices.

### 4.7 Exercises

1. In 4.1 - 4.8 describe $-A$ and 0.

2. Let $A$ be an $n \times n$ matrix. Show $A$ equals the sum of a symmetric and a skew symmetric matrix.

3. Show every skew symmetric matrix has all zeros down the main diagonal. The main diagonal consists of every entry of the matrix which is of the form $a_{ii}$. It runs from the upper left down to the lower right.

4. We used the fact that the columns of a matrix $A$ are independent if and only if $Ax = 0$ has only the zero solution for $x$. Why is this so?

5. If $A$ is $m \times n$ where $n > m$, explain why there exists $x \in \mathbb{R}^n$ such that $Ax = 0$ but $x \neq 0$.

6. Using only the properties 4.1 - 4.8 show $-A$ is unique.

7. Using only the properties 4.1 - 4.8 show 0 is unique.

8. Using only the properties 4.1 - 4.8 show $0A = 0$. Here the 0 on the left is the scalar 0 and the 0 on the right is the zero for $m \times n$ matrices.

9. Using only the properties 4.1 - 4.8 and previous problems show $(-1)A = -A$.

10. Prove that $I_mA = A$ where $A$ is an $m \times n$ matrix.

11. Let $A$ and be a real $m \times n$ matrix and let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Show $(Ax,y)_{\mathbb{R}^m} = (x,A^Ty)_{\mathbb{R}^n}$ where $(\cdot, \cdot)_{\mathbb{R}^k}$ denotes the dot product in $\mathbb{R}^k$. You need to know about the dot product. It will be discussed later but hopefully it has been seen in physics or calculus.

12. Use the result of Problem 11 to verify directly that $(AB)^T = B^TA^T$ without making any reference to subscripts. However, note that the treatment in the chapter did not depend on a dot product.

13. Let $x = (-1, -1, 1)$ and $y = (0, 1, 2)$. Find $x^Ty$ and $xy^T$ if possible.

14. Give an example of matrices, $A, B, C$ such that $B \neq C$, $A \neq 0$, and yet $AB = AC$. 
15. Let \( A = \begin{pmatrix} 1 & 1 \\ -2 & -1 \\ 1 & 2 \end{pmatrix} \), \( B = \begin{pmatrix} 1 & -1 & -2 \\ 2 & 1 & -2 \end{pmatrix} \), and \( C = \begin{pmatrix} 1 & 1 & -3 \\ -1 & 2 & 0 \\ -3 & -1 & 0 \end{pmatrix} \). Find if possible the following products. \( AB, BA, AC, CA, CB, BC \).

16. Show \((AB)^{-1} = B^{-1}A^{-1}\).

17. Show that if \( A \) is an invertible \( n \times n \) matrix, then so is \( A^T \) and \((A^T)^{-1} = (A^{-1})^T\).

18. Show that if \( A \) is an \( n \times n \) invertible matrix and \( x \) is a \( n \times 1 \) matrix such that \( Ax = b \) for \( b \) an \( n \times 1 \) matrix, then \( x = A^{-1}b \).

19. Give an example of a matrix \( A \) such that \( A^2 = I \) and yet \( A \neq I \) and \( A \neq -I \).

20. Give an example of matrices, \( A, B \) such that neither \( A \) nor \( B \) equals zero and yet \( AB = 0 \).

21. Give another example other than the one given in this section of two square matrices, \( A \) and \( B \) such that \( AB \neq BA \).

22. Suppose \( A \) and \( B \) are square matrices of the same size. Which of the following are correct?
   
   \[ (a) \ (A - B)^2 = A^2 - 2AB + B^2 \]
   
   \[ (b) \ (AB)^2 = A^2B^2 \]
   
   \[ (c) \ (A + B)^2 = A^2 + 2AB + B^2 \]
   
   \[ (d) \ (A + B)^2 = A^2 + AB + BA + B^2 \]
   
   \[ (e) \ A^2B^2 = A (AB) B \]
   
   \[ (f) \ (A + B)^3 = A^3 + 3A^2B + 3AB^2 + B^3 \]
   
   \[ (g) \ (A + B) (A - B) = A^2 - B^2 \]
   
   \[ (h) \ None \ of \ the \ above. \ They \ are \ all \ wrong. \]
   
   \[ (i) \ All \ of \ the \ above. \ They \ are \ all \ right. \]

23. Let \( A = \begin{pmatrix} -1 & -1 \\ 3 & 3 \end{pmatrix} \). Find all \( 2 \times 2 \) matrices, \( B \) such that \( AB = 0 \).

24. Prove that if \( A^{-1} \) exists and \( Ax = 0 \) then \( x = 0 \).

25. Let \( A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 1 & 0 & 2 \end{pmatrix} \).

   Find \( A^{-1} \) if possible. If \( A^{-1} \) does not exist, determine why.

26. Let \( A = \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \).

   Find \( A^{-1} \) if possible. If \( A^{-1} \) does not exist, determine why.

27. Let \( A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 4 & 5 & 10 \end{pmatrix} \).

   Find \( A^{-1} \) if possible. If \( A^{-1} \) does not exist, determine why.
28. Let

\[ A = \begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 1 & 2 & 0 \\ 2 & 1 & -3 & 2 \\ 1 & 2 & 1 & 2 \end{pmatrix} \]

Find \( A^{-1} \) if possible. If \( A^{-1} \) does not exist, determine why.

29. Let

\[ A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \]

Find \( A^{-1} \) if possible. If \( A^{-1} \) does not exist, determine why. Do this in \( \mathbb{Q}^2 \) and in \( \mathbb{Z}_5^2 \).

30. Let

\[ A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \]

Find \( A^{-1} \) if possible. If \( A^{-1} \) does not exist, determine why. Do this in \( \mathbb{Q}^2 \) and in \( \mathbb{Z}_5^2 \).

31. If you have any system of equations \( Ax = b \), let \( \ker(A) = \{ x : Ax = 0 \} \). Show that all solutions of the system \( Ax = b \) are in \( \ker(A) + y_p \) where \( Ay_p = b \). This means that every solution of this last equation is of the form \( y_p + z \) where \( Az = 0 \).

32. Write the solution set of the following system as the span of vectors and find a basis for the solution space of the following system.

\[ \begin{pmatrix} 1 & -1 & 2 \\ 1 & -2 & 1 \\ 3 & -4 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \]

33. Using Problem 32 find the general solution to the following linear system.

\[ \begin{pmatrix} 1 & -1 & 2 \\ 1 & -2 & 1 \\ 3 & -4 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix} \]

34. Write the solution set of the following system as the span of vectors and find a basis for the solution space of the following system.

\[ \begin{pmatrix} 0 & -1 & 2 \\ 1 & -2 & 1 \\ 1 & -4 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \]

35. Using Problem 34 find the general solution to the following linear system.

\[ \begin{pmatrix} 0 & -1 & 2 \\ 1 & -2 & 1 \\ 1 & -4 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \]

36. Write the solution set of the following system as the span of vectors and find a basis for the solution space of the following system.

\[ \begin{pmatrix} 1 & -1 & 2 \\ 1 & -2 & 0 \\ 3 & -4 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \]
37. Using Problem 36 find the general solution to the following linear system.

\[
\begin{pmatrix}
1 & -1 & 2 \\
1 & -2 & 0 \\
3 & -4 & 4
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= 
\begin{pmatrix}
1 \\
2 \\
4
\end{pmatrix}.
\]

38. Show that 4.24 is valid for \( p = -1 \) if and only if each block has an inverse and that this condition holds if and only if \( A \) is invertible.

39. Let \( A \) be an \( n \times n \) matrix and let \( P^{ij} \) be the permutation matrix which switches the \( i^{th} \) and \( j^{th} \) rows of the identity. Show that \( P^{ij}AP^{ij} \) produces a matrix which is similar to \( A \) which switches the \( i^{th} \) and \( j^{th} \) entries on the main diagonal.

40. You could define column operations by analogy to row operations. That is, you switch two columns, multiply a column by a nonzero scalar, or add a scalar multiple of a column to another column. Let \( E \) be one of these column operations applied to the identity matrix. Show that \( AE \) produces the column operation on \( A \) which was used to define \( E \).

41. Consider the symmetric 3 \( \times \) 3 matrices, those for which \( A = A^T \). Show that with respect to the usual notions of addition and scalar multiplication this is a vector space of dimension 6. What is the dimension of the set of skew symmetric matrices?
Chapter 5

Linear Transformations

This chapter is on functions which map a vector space to another one which are also linear. The description of these is in the following definition. Linear algebra is all about understanding these kinds of mappings.

**Definition 5.0.1** Let $V$ and $W$ be two finite dimensional vector spaces. A function, $L$ which maps $V$ to $W$ is called a linear transformation and written $L \in \mathcal{L}(V,W)$ if for all scalars $\alpha$ and $\beta$, and vectors $v, w$, 

$$L(\alpha v + \beta w) = \alpha L(v) + \beta L(w).$$

**Example 5.0.2** Let $V = \mathbb{R}^3, W = \mathbb{R}$, and let $a \in \mathbb{R}^3$ be given vector in $V$. Define $T : V \to W$

$$Tv \equiv \sum_{i=1}^{3} a_i v_i$$

It is left as an exercise to verify that this is indeed linear. In fact, it is just the dot product for those who remember calculus. $Tv \equiv a \cdot v$. In fact, this is similar to the general case for vector spaces $\mathbb{F}^n = V, \mathbb{F}^m = W$ in a way to be explained later. Here is an interesting observation.

**Proposition 5.0.3** Let $L : \mathbb{F}^n \to \mathbb{F}^m$ be linear. Then there exists a unique $m \times n$ matrix $A$ such that

$$Lx = Ax$$

for all $x$. Also, matrix multiplication yields a linear transformation.

**Proof:** Note that

$$x = \sum_{i=1}^{n} x_i e_i$$

and so

$$Lx = L \left( \sum_{i=1}^{n} x_i e_i \right) = \sum_{i=1}^{n} x_i Le_i = \begin{pmatrix} Le_1 & \cdots & Le_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} Le_1 & \cdots & Le_n \end{pmatrix} x$$

The matrix is $A$. The last claim follows from the properties of matrix multiplication. \(\blacksquare\)

I will abuse terminology slightly and say that a $m \times n$ matrix is one to one if the linear transformation it determines is one to one, similarly for the term onto.
### 5.1 $\mathcal{L}(V, W)$ As A Vector Space

The linear transformations can be considered as a vector space as described next.

**Definition 5.1.1** Given $L, M \in \mathcal{L}(V, W)$ define a new element of $\mathcal{L}(V, W)$, denoted by $L + M$ according to the rule

$$(L + M)v \equiv Lv + Mv.$$

For $\alpha$ a scalar and $L \in \mathcal{L}(V, W)$, define $\alpha L \in \mathcal{L}(V, W)$ by

$$\alpha L(v) \equiv \alpha (Lv).$$

You should verify that all the axioms of a vector space hold for $\mathcal{L}(V, W)$ with the above definitions of vector addition and scalar multiplication. In fact, is just a subspace of the set of functions mapping $V$ to $W$ which is a vector space thanks to Example 3.0.3. What about the dimension of $\mathcal{L}(V, W)$? What about a basis for $\mathcal{L}(V, W)$?

Before answering this question, here is a useful lemma. It gives a way to define linear transformations and a way to tell when two of them are equal.

**Lemma 5.1.2** Let $V$ and $W$ be vector spaces and suppose $\{v_1, \ldots, v_n\}$ is a basis for $V$. Then if $L : V \to W$ is given by $Lv_k = w_k \in W$ and

$$L \left( \sum_{k=1}^{n} a_k v_k \right) = \sum_{k=1}^{n} a_k Lv_k = \sum_{k=1}^{n} a_k w_k$$

then $L$ is well defined and is in $\mathcal{L}(V, W)$. Also, if $L, M$ are two linear transformations such that $Lv_k = Mv_k$ for all $k$, then $M = L$.

**Proof:** $L$ is well defined on $V$ because, since $\{v_1, \ldots, v_n\}$ is a basis, there is exactly one way to write a given vector of $V$ as a linear combination. Next, observe that $L$ is obviously linear from the definition. If $L, M$ are equal on the basis, then if $\sum_{k=1}^{n} a_k v_k$ is an arbitrary vector of $V$,

$$L \left( \sum_{k=1}^{n} a_k v_k \right) = \sum_{k=1}^{n} a_k Lv_k = \sum_{k=1}^{n} a_k Mv_k = M \left( \sum_{k=1}^{n} a_k v_k \right)$$

and so $L = M$ because they give the same result for every vector in $V$. ■

The message is that when you define a linear transformation, it suffices to tell what it does to a basis.

**Example 5.1.3** A basis for $\mathbb{R}^2$ is

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Suppose $T$ is a linear transformation which satisfies

$$T \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \ T \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

Find $T \begin{pmatrix} 3 \\ 2 \end{pmatrix}$. 

---

1Note that this is the standard way of defining the sum of two functions.
\[
T \begin{pmatrix} 3 \\ 2 \end{pmatrix} = T \left( 2 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right) = 2T \begin{pmatrix} 1 \\ 1 \end{pmatrix} + T \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 2 \begin{pmatrix} 2 \\ 1 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}
\]

**Theorem 5.1.4** Let \( V \) and \( W \) be finite dimensional linear spaces of dimension \( n \) and \( m \) respectively. Then \( \dim(\mathcal{L}(V, W)) = mn \).

**Proof:** Let two sets of bases be 

\[ \{v_1, \ldots, v_n\} \text{ and } \{w_1, \ldots, w_m\} \]

for \( V \) and \( W \) respectively. Using Lemma 5.1.2, let \( w_i v_j \in \mathcal{L}(V, W) \) be the linear transformation defined on the basis, \( \{v_1, \ldots, v_n\} \), by

\[ w_i v_k (v_j) \equiv w_i \delta_{jk} \]

where \( \delta_{ik} = 1 \) if \( i = k \) and 0 if \( i \neq k \). I will show that \( \mathcal{L} \subseteq \mathcal{L}(V, W) \) is a linear combination of these special linear transformations called dyadics, also rank one transformations.

Then let \( L \in \mathcal{L}(V, W) \). Since \( \{w_1, \ldots, w_m\} \) is a basis, there exist constants, \( d_{jk} \) such that

\[ L v_r = \sum_{j=1}^{m} d_{jr} w_j \]

Now consider the following sum of dyadics.

\[ \sum_{j=1}^{m} \sum_{i=1}^{n} d_{ji} w_j v_i \]

Apply this to \( v_r \). This yields

\[ \sum_{j=1}^{m} \sum_{i=1}^{n} d_{ji} w_j v_i (v_r) = \sum_{j=1}^{m} \sum_{i=1}^{n} d_{ji} w_j \delta_{ir} = \sum_{j=1}^{m} d_{jr} w_j = L v_r \] (5.1)

Therefore, \( L = \sum_{j=1}^{m} \sum_{i=1}^{n} d_{ji} w_j v_i \) showing the span of the dyadics is all of \( \mathcal{L}(V, W) \).

Now consider whether these dyadics form a linearly independent set. Suppose

\[ \sum_{i,k} d_{ik} w_i v_k = 0. \]

Are all the scalars \( d_{ik} \) equal to 0?

\[ 0 = \sum_{i,k} d_{ik} w_i v_k (v_l) = \sum_{i=1}^{m} d_{il} w_i \]

and so, since \( \{w_1, \ldots, w_m\} \) is a basis, \( d_{il} = 0 \) for each \( i = 1, \ldots, m \). Since \( l \) is arbitrary, this shows \( d_{il} = 0 \) for all \( i \) and \( l \). Thus these linear transformations form a basis and this shows that the dimension of \( \mathcal{L}(V, W) \) is \( mn \) as claimed because there are \( m \) choices for the \( w_i \) and \( n \) choices for the \( v_j \).

Note that from 5.1.1 these coefficients which obtain \( L \) as a linear combination of the diadics are given by the equation

\[ \sum_{j=1}^{m} d_{jr} w_j = L v_r \] (5.2)

Thus \( L v_r \) is in the span of the \( w_j \) and the scalars in the linear combination are \( d_{1r}, d_{2r}, \ldots, d_{mr} \).
5.2 The Matrix Of A Linear Transformation

In order to do computations based on a linear transformation, we usually work with its matrix. This is what is described here.

Theorem 5.1.4 says that the rank one transformations defined there in terms of two bases, one for $V$ and the other for $W$ are a basis for $L(V,W)$. Thus if $A \in L(V,W)$, there are scalars $A_{ij}$ such that

$$A = \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} w_{i} v_{j}$$

Here we have $1 \leq i \leq n$ and $1 \leq j \leq m$. We can arrange these scalars in a rectangular shape as follows.

$$
\begin{pmatrix}
A_{11} & A_{12} & \cdots & A_{1(n-1)} & A_{1n} \\
A_{21} & A_{22} & \cdots & A_{2(n-1)} & A_{2n} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
A_{m1} & A_{m2} & \cdots & A_{m(n-1)} & A_{mn}
\end{pmatrix}
$$

Here this is an $m \times n$ matrix because it has $m$ rows and $n$ columns. It is called the matrix of the linear transformation $A$ with respect to the two bases $\{v_1, \cdots, v_n\}$ for $V$ and $\{w_1, \cdots, w_m\}$ for $W$.

Now, as noted earlier, if $v = \sum_{r=1}^{n} x_{r} v_{r}$,

$$Av = \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} w_{i} v_{j} \left( \sum_{r=1}^{n} x_{r} v_{r} \right)$$

$$= \sum_{r=1}^{n} x_{r} \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} w_{i} v_{j} (v_{r}) = \sum_{r=1}^{n} x_{r} \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} w_{i} \delta_{jr}$$

$$= \sum_{r=1}^{n} x_{r} \sum_{i=1}^{n} A_{ir} w_{i} = \sum_{r=1}^{n} \left( \sum_{i=1}^{n} A_{ir} x_{r} \right) w_{i}$$

What does this show? It shows that if the component vector of $v$ is

$$\begin{pmatrix}
x_{1} \\
\vdots \\
x_{n}
\end{pmatrix}$$

meaning that $v = \sum_{i} x_{i} v_{i}$, then the component vector of $w$ has $i^{th}$ component equal to

$$\sum_{r=1}^{n} A_{ir} x_{r}$$

This motivates the definition of what it means to multiply matrices.

**Definition 5.2.1** Let $A$ be an $m \times n$ matrix and let $x \in \mathbb{F}^{n}$. Then $Ax \in \mathbb{F}^{m}$ and if $Ax = y$, then

$$\begin{pmatrix}
y_{1} \\
\vdots \\
y_{m}
\end{pmatrix} = \begin{pmatrix}
\sum_{j} A_{1j} x_{j} \\
\vdots \\
\sum_{j} A_{mj} x_{j}
\end{pmatrix}$$

In short,

$$(Ax)_{i} = \sum_{j} A_{ij} x_{j}$$
The idea is that acting on a vector $v$ with a linear transformation $T$ yields a new vector $w$ whose component vector is obtained as the matrix of the linear transformation times the component vector of $v$. It is helpful for some of us to think of this in terms of diagrams. On the other hand, some people hate such diagrams. Use them if it helps. Otherwise ignore them and go right to the algebraic definition.

Let $\beta = \{v_1, \cdots, v_n\}$ be a basis for $V$ and let $\{w_1, \cdots, w_m\} = \gamma$ be a basis for $W$. Then let $q_\beta : \mathbb{F}^n \rightarrow V$, $q_\gamma : \mathbb{F}^m \rightarrow W$ be defined as

$$q_\beta x = \sum_{i=1}^{n} x_i v_i, \quad q_\gamma y = \sum_{j=1}^{m} y_j w_j$$

Thus these mappings are linear and take the component vector to the vector determined by the component vector.

Then the diagram which describes the matrix of the linear transformation $L$ is in the following picture.

$$\begin{array}{ccc}
\beta = \{v_1, \cdots, v_n\} & V & W \{w_1, \cdots, w_m\} = \gamma \\
q_\beta \uparrow & \circ & \uparrow q_\gamma \\
\mathbb{F}^n & \rightarrow & \mathbb{F}^m \\
[L]_{\gamma\beta}
\end{array}$$

In terms of this diagram, the matrix $[L]_{\gamma\beta}$ is the matrix chosen to make the diagram “commute”. It is the matrix of the linear transformation because it takes the component vector of $v$ to the component vector for $Lv$. As implied by the diagram and as shown above, for $A = [L]_{\gamma\beta}$,

$$Lv_i = \sum_{j=1}^{m} A_{ji} w_j$$

It may be useful to write this in the form

$$\begin{pmatrix}
Lv_1 \\
\vdots \\
Lv_n
\end{pmatrix} = 
\begin{pmatrix}
w_1 \\
\vdots \\
w_m
\end{pmatrix} A, \ A \text{ is } m \times n$$

and multiply formally as if the $Lv_i, w_j$ were numbers.

**Example 5.2.2** Let $L \in \mathcal{L} (\mathbb{F}^n, \mathbb{F}^m)$ and let the two bases be $\left\{ e_1, \cdots, e_n \right\}$, $\left\{ e_1, \cdots, e_m \right\}$, $e_i$ denoting the column vector of zeros except for a 1 in the $i^{th}$ position. Then from the above, you need to have

$$Le_i = \sum_{j=1}^{m} A_{ji} e_j$$

which says that

$$\begin{pmatrix}
Le_1 \\
\vdots \\
Le_n
\end{pmatrix}_{m \times n} = 
\begin{pmatrix}
e_1 \\
\vdots \\
e_m
\end{pmatrix}_{m \times m} A_{m \times n}$$

and so $Le_i$ equals the $i^{th}$ column of $A$. In other words,

$$A = \begin{pmatrix}
Le_1 \\
\vdots \\
Le_n
\end{pmatrix}.$$
What is the matrix of this linear transformation with respect to this basis? Using \([5.4]\),
\[
\begin{pmatrix}
0 & 1 & 2x & 3x^2
\end{pmatrix} = \begin{pmatrix}
1 & x & x^2
\end{pmatrix} \begin{bmatrix} D \end{bmatrix}_{\gamma\beta}.
\]
It follows from this that the first column of \([D]_{\gamma\beta}\) is
\[
\begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}
\]
The next three columns of \([D]_{\gamma\beta}\) are
\[
\begin{pmatrix}
1 \\
0 \\
0
\end{pmatrix}, \begin{pmatrix}
0 \\
2 \\
0
\end{pmatrix}, \begin{pmatrix}
0 \\
0 \\
3
\end{pmatrix}
\]
and so
\[
[D]_{\gamma\beta} = \begin{pmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 2 & 0 \\
0 & 0 & 0 & 3
\end{pmatrix}.
\]
Say you have \(a + bx + cx^2 + dx^3\). Then doing \(D\) to it gives \(b + 2cx + 3dx^2\). The component vector of the function is
\[
\begin{pmatrix}
a & b & c & d
\end{pmatrix}^T
\]
and after doing \(D\) to the function, you get for the component vector
\[
\begin{pmatrix}
b & 2c & 3d
\end{pmatrix}^T
\]
This is the same result you get when you multiply by \([D]\).
\[
\begin{pmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 2 & 0 \\
0 & 0 & 0 & 3
\end{pmatrix} \begin{pmatrix}
a \\
b \\
c \\
d
\end{pmatrix} = \begin{pmatrix}
b \\
2c \\
3d
\end{pmatrix}
\]
Of course, this is what it means to be the matrix of the transformation.

Now consider the important case where \(V = \mathbb{F}^n\), \(W = \mathbb{F}^m\), and the basis chosen is the standard basis of vectors \(e_i\) described above.
\[
\beta = \{e_1, \ldots, e_n\}, \quad \gamma = \{e_1, \ldots, e_m\}
\]
Let \(L\) be a linear transformation from \(\mathbb{F}^n\) to \(\mathbb{F}^m\) and let \(A\) be the matrix of the transformation with respect to these bases. In this case the coordinate maps \(q_\beta\) and \(q_\gamma\) are simply the identity maps on \(\mathbb{F}^n\) and \(\mathbb{F}^m\) respectively, and can be accomplished by simply multiplying by the appropriate sized identity matrix. The requirement that \(A\) is the matrix of the transformation amounts to
\[
Lb = Ab
\]
What about the situation where different pairs of bases are chosen for \(V\) and \(W\)? How are the two matrices with respect to these choices related? Consider the following diagram which illustrates the situation.
In this diagram \( q_{\beta_i} \) and \( q_{\gamma_i} \) are coordinate maps as described above. From the diagram,
\[
q_{\gamma_1}^{-1} q_{\gamma_2} A_2 q_{\beta_1}^{-1} q_{\beta_2} = A_1,
\]
where \( q_{\beta_i}^{-1} \) and \( q_{\gamma_i}^{-1} \) are one to one, onto, and linear maps which may be accomplished by multiplication by a square matrix. Thus there exist matrices \( P, Q \) such that
\[
P A_2 Q = A_1.
\]

**Example 5.2.4** Let \( \beta = \{v_1, \ldots, v_n\} \) and \( \gamma = \{w_1, \ldots, w_n\} \) be two bases for \( V \). Let \( L \) be the linear transformation which maps \( v_i \) to \( w_i \). Find \( [L]_{\gamma\beta} \).

Letting \( \delta_{ij} \) be the symbol which equals 1 if \( i = j \) and 0 if \( i \neq j \), it follows that \( L = \sum_{i,j} \delta_{ij} w_i v_j \) and so \( [L]_{\gamma\beta} = I \) the identity matrix.

**Definition 5.2.5** In the special case where \( V = W \) and only one basis is used for \( V = W \), this becomes
\[
q_{\beta_1}^{-1} q_{\beta_2} A_2 q_{\beta_1}^{-1} q_{\beta_2} = A_1.
\]

Letting \( S \) be the matrix of the linear transformation \( q_{\beta_2}^{-1} q_{\beta_1} \) with respect to the standard basis vectors in \( \mathbb{F}^n \),
\[
S^{-1} A_2 S = A_1. \tag{5.5}
\]
When this occurs, \( A_1 \) is said to be similar to \( A_2 \) and \( A_1 \rightarrow S^{-1} A S \) is called a similarity transformation.

Recall the following.

**Definition 5.2.6** Let \( S \) be a set. The symbol \( \sim \) is called an equivalence relation on \( S \) if it satisfies the following axioms.

1. \( x \sim x \) for all \( x \in S \). (Reflexive)
2. If \( x \sim y \) then \( y \sim x \). (Symmetric)
3. If \( x \sim y \) and \( y \sim z \), then \( x \sim z \). (Transitive)

**Definition 5.2.7** \([x]\) denotes the set of all elements of \( S \) which are equivalent to \( x \) and \([x]\) is called the equivalence class determined by \( x \) or just the equivalence class of \( x \).

Also recall the notion of equivalence classes.

**Theorem 5.2.8** Let \( \sim \) be an equivalence class defined on a set \( S \) and let \( \mathcal{H} \) denote the set of equivalence classes. Then if \([x]\) and \([y]\) are two of these equivalence classes, either \( x \sim y \) and \([x] = [y]\) or it is not true that \( x \sim y \) and \([x] \cap [y] = \emptyset \).

**Theorem 5.2.9** In the vector space of \( n \times n \) matrices, define
\[
A \sim B
\]
if there exists an invertible matrix \( S \) such that
\[
A = S^{-1} BS.
\]
Then \( \sim \) is an equivalence relation and \( A \sim B \) if and only if whenever \( V \) is an \( n \) dimensional vector space, there exists \( L \in \mathcal{L}(V,V) \) and bases \( \{v_1, \ldots, v_n\} \) and \( \{w_1, \ldots, w_n\} \) such that \( A \) is the matrix of \( L \) with respect to \( \{v_1, \ldots, v_n\} \) and \( B \) is the matrix of \( L \) with respect to \( \{w_1, \ldots, w_n\} \).
**Proof:** \( A \sim A \) because \( S = I \) works in the definition. If \( A \sim B \), then \( B \sim A \), because

\[
A = S^{-1}BS
\]
implies \( B = SAS^{-1} \). If \( A \sim B \) and \( B \sim C \), then

\[
A = S^{-1}BS, \quad B = T^{-1}CT
\]
and so

\[
A = S^{-1}T^{-1}CTS = (TS)^{-1}CTS
\]
which implies \( A \sim C \). This verifies the first part of the conclusion.

Now let \( V \) be an \( n \) dimensional vector space, \( A \sim B \) so \( A = S^{-1}BS \) and pick a basis for \( V \),

\[
\beta \equiv \{ v_1, \cdots, v_n \}
\]
Define \( L \in \mathcal{L}(V, V) \) by

\[
Lv_i = \sum_j a_{ji}v_j
\]
where \( A = (a_{ij}) \). Thus \( A \) is the matrix of the linear transformation \( L \). Consider the diagram

\[
\begin{array}{ccc}
\mathbb{F}^n & B & \mathbb{F}^n \\
q_\gamma \downarrow & \circ & q_\gamma \downarrow \\
V & L & V \\
q_\beta \uparrow & \circ & q_\beta \uparrow \\
\mathbb{F}^n & A & \mathbb{F}^n
\end{array}
\]
where \( q_\gamma \) is chosen to make the diagram commute. Thus we need \( S = q_\gamma^{-1}q_\beta \) which requires

\[
q_\gamma = q_\beta S^{-1}
\]
Then it follows that \( B \) is the matrix of \( L \) with respect to the basis

\[
\{ q_\gamma e_1, \cdots, q_\gamma e_n \} \equiv \{ w_1, \cdots, w_n \}.
\]
That is, \( A \) and \( B \) are matrices of the same linear transformation \( L \). Conversely, suppose whenever \( V \) is an \( n \) dimensional vector space, there exists \( L \in \mathcal{L}(V, V) \) and bases \( \{ v_1, \cdots, v_n \} \) and \( \{ w_1, \cdots, w_n \} \) such that \( A \) is the matrix of \( L \) with respect to \( \{ v_1, \cdots, v_n \} \) and \( B \) is the matrix of \( L \) with respect to \( \{ w_1, \cdots, w_n \} \). Then it was shown above that \( A \sim B \). □

What if the linear transformation consists of multiplication by a matrix \( A \) and you want to find the matrix of this linear transformation with respect to another basis? Is there an easy way to do it? The next proposition considers this.

**Proposition 5.2.10** Let \( A \) be an \( m \times n \) matrix and consider it as a linear transformation by multiplication on the left by \( A \). Then the matrix \( M \) of this linear transformation with respect to the bases \( \beta = \{ u_1, \cdots, u_n \} \) for \( \mathbb{F}^n \) and \( \gamma = \{ w_1, \cdots, w_m \} \) for \( \mathbb{F}^m \) is given by

\[
M = \left( \begin{array}{c}
w_1 \\
\vdots \\
w_m
\end{array} \right)^{-1} A \left( \begin{array}{c}
u_1 \\
\vdots \\
u_n
\end{array} \right)
\]
where \( \left( \begin{array}{c}
w_1 \\
\vdots \\
w_m
\end{array} \right) \) is the \( m \times m \) matrix which has \( w_j \) as its \( j^{th} \) column. Note that also

\[
\left( \begin{array}{c}
w_1 \\
\vdots \\
w_m
\end{array} \right) M \left( \begin{array}{c}
u_1 \\
\vdots \\
u_n
\end{array} \right)^{-1} = A
\]
5.2. THE MATRIX OF A LINEAR TRANSFORMATION

**Proof:** Consider the following diagram.

\[
\begin{array}{ccc}
A & F^n & \rightarrow & F^m \\
\uparrow q_\beta & & & \uparrow q_\gamma \\
F^n & \rightarrow & F^m \\
M & & \\
\end{array}
\]

Here the coordinate maps are defined in the usual way. Thus

\[q_\beta(x) = \sum_{i=1}^{n} x_i u_i = \left( \begin{array}{c} u_1 \\ \vdots \\ u_n \end{array} \right) x\]

Therefore, \(q_\beta\) can be considered the same as multiplication of a vector in \(F^n\) on the left by the matrix \(\left( \begin{array}{c} u_1 \\ \vdots \\ u_n \end{array} \right)\). Similar considerations apply to \(q_\gamma\). Thus it is desired to have the following for an arbitrary \(x \in F^m\).

\[A \left( \begin{array}{c} u_1 \\ \vdots \\ u_n \end{array} \right) x = \left( \begin{array}{c} w_1 \\ \vdots \\ w_n \end{array} \right) Mx\]

Therefore, the conclusion of the proposition follows. \(\blacksquare\)

The second formula in the above is pretty useful. You might know the matrix \(M\) of a linear transformation with respect to a funny basis and this formula gives the matrix of the linear transformation in terms of the usual basis which is really what you want.

**Definition 5.2.11** Let \(A \in \mathcal{L}(X,Y)\) where \(X\) and \(Y\) are finite dimensional vector spaces. Define \(\text{rank}(A)\) to equal the dimension of \(A(X)\).

**Lemma 5.2.12** Let \(M\) be an \(m \times n\) matrix. Then \(M\) can be considered as a linear transformation as follows.

\[M(x) = Mx\]

That is, you multiply on the left by \(M\).

**Proof:** This follows from the properties of matrix multiplication. In particular,

\[M(ax + by) = aMx + bMy\]  \(\blacksquare\)

Note also that, as explained earlier, the image of this transformation is just the span of the columns, known as the column space.

The following theorem explains how the rank of \(A\) is related to the rank of the matrix of \(A\).

**Theorem 5.2.13** Let \(A \in \mathcal{L}(X,Y)\). Then \(\text{rank}(A) = \text{rank}(M)\) where \(M\) is the matrix of \(A\) taken with respect to a pair of bases for the vector spaces \(X\) and \(Y\). Here \(M\) is considered as a linear transformation by matrix multiplication.

**Proof:** Recall the diagram which describes what is meant by the matrix of \(A\). Here the two bases are as indicated.

\[\beta = \{v_1, \ldots, v_n\} \quad X \quad A \quad Y \quad \{w_1, \ldots, w_m\} = \gamma\]

\[q_\beta \uparrow \quad \uparrow \quad q_\gamma \]

\[F^n \quad M \quad F^m\]

Let \(\{Ax_1, \ldots, Ax_r\}\) be a basis for \(AX\) Then

\[\left\{ q_\gamma Mq_\beta^{-1}x_1, \ldots, q_\gamma Mq_\beta^{-1}x_r \right\}\]

is a basis for \(AX\). It follows that

\[\left\{ Mq_X^{-1}x_1, \ldots, Mq_X^{-1}x_r \right\}\]

is linearly independent and so \(\text{rank}(A) \leq \text{rank}(M)\). However, one could interchange the roles of \(M\) and \(A\) in the above argument and thereby turn the inequality around. \(\blacksquare\)

The following result is a summary of many concepts.
Theorem 5.2.14 Let \( L \in \mathcal{L}(V, V) \) where \( V \) is a finite dimensional vector space. Then the following are equivalent.

1. \( L \) is one to one.
2. \( L \) maps a basis to a basis.
3. \( L \) is onto.
4. If \( Lv = 0 \) then \( v = 0 \).

Proof: Suppose first \( L \) is one to one and let \( \beta = \{v_i\}_{i=1}^n \) be a basis. Then if \( \sum_{i=1}^n c_iLv_i = 0 \) it follows \( L(\sum_{i=1}^n c_i v_i) = 0 \) which means that since \( L(0) = 0 \), and \( L \) is one to one, it must be the case that \( \sum_{i=1}^n c_i v_i = 0 \). Since \( \{v_i\} \) is a basis, each \( c_i = 0 \) which shows \( \{Lv_i\} \) is a linearly independent set. Since there are \( n \) of these, it must be that this is a basis.

Now suppose 2.). Then letting \( \{v_i\} \) be a basis, and \( y \in V \), it follows from part 2.) that there are constants, \( \{c_i\} \) such that \( y = \sum_{i=1}^n c_iLv_i = L(\sum_{i=1}^n c_i v_i) \). Thus \( L \) is onto. It has been shown that 2.) implies 3.).

Now suppose 3.). Then \( L(V) = V \). If \( \{v_1, \ldots, v_n\} \) is a basis of \( V \), then \( V = \text{span}(Lv_1, \ldots, Lv_n) \). It follows that \( \{Lv_1, \ldots, Lv_n\} \) must be linearly independent because if not, one of the vectors could be deleted and you would then have a spanning set with fewer vectors than \( \text{dim}(V) \). If \( Lv = 0 \),

\[
\begin{align*}
v &= \sum_i x_i v_i \\
0 &= \sum_i x_i Lv_i
\end{align*}
\]

then doing \( L \) to both sides,

\[
\begin{align*}
0 &= \sum_i x_i Lv_i
\end{align*}
\]

which implies each \( x_i = 0 \) and consequently \( v = 0 \). Thus 4. follows.

Now suppose 4.) and suppose \( Lv = Lw \). Then \( L(v - w) = 0 \) and so by 4.), \( v - w = 0 \) showing that \( L \) is one to one.

Also it is important to note that composition of linear transformations corresponds to multiplication of the matrices. Consider the following diagram in which \( [A]_{\gamma\beta} \) denotes the matrix of \( A \) relative to the bases \( \gamma \) on \( Y \) and \( \beta \) on \( X \), \([B]_{\delta\gamma}\) defined similarly.

\[
\begin{array}{cccccc}
X & A & Y & B & Z \\
\uparrow q_\beta & \circ & \uparrow q_\gamma & \circ & \uparrow q_\delta \\
F^n & [A]_{\gamma\beta} & F^m & [B]_{\delta\gamma} & F^p
\end{array}
\]

where \( A \) and \( B \) are two linear transformations, \( A \in \mathcal{L}(X, Y) \) and \( B \in \mathcal{L}(Y, Z) \). Then \( B \circ A \in \mathcal{L}(X, Z) \) and so it has a matrix with respect to bases given on \( X \) and \( Z \), the coordinate maps for these bases being \( q_\beta \) and \( q_\delta \) respectively. Then

\[
B \circ A = q_\delta [B]_{\delta\gamma} q_\gamma^{-1} q_\gamma [A]_{\gamma\beta} q_\beta^{-1} = q_\delta [B]_{\delta\gamma} [A]_{\gamma\beta} q_\beta^{-1}.
\]

But this shows that \( [B]_{\delta\gamma} [A]_{\gamma\beta} \) plays the role of \( [B \circ A]_{\delta\beta} \), the matrix of \( B \circ A \). Hence the matrix of \( B \circ A \) equals the product of the two matrices \( [A]_{\gamma\beta} \) and \( [B]_{\delta\gamma} \). Of course it is interesting to note that although \( [B \circ A]_{\delta\beta} \) must be unique, the matrices, \( [A]_{\gamma\beta} \) and \( [B]_{\delta\gamma} \) are not unique because they depend on \( \gamma \), the basis chosen for \( Y \).

Theorem 5.2.15 The matrix of the composition of linear transformations equals the product of the matrices of these linear transformations.
5.3 Rotations About A Given Vector

As an application, consider the problem of rotating counter clockwise about a given unit vector which is possibly not one of the unit vectors in coordinate directions. First consider a pair of perpendicular unit vectors, \( \mathbf{u}_1 \) and \( \mathbf{u}_2 \) and the problem of rotating in the counterclockwise direction about \( \mathbf{u}_3 \) where \( \mathbf{u}_3 = \mathbf{u}_1 \times \mathbf{u}_2 \) so that \( \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3 \) forms a right handed orthogonal coordinate system. See the appendix on the cross product if this is not familiar. Thus the vector \( \mathbf{u}_3 \) is coming out of the page.

Let \( T \) denote the desired rotation. Then
\[
T(\alpha \mathbf{u}_1 + \beta \mathbf{u}_2 + \gamma \mathbf{u}_3) = \alpha T \mathbf{u}_1 + \beta T \mathbf{u}_2 + \gamma T \mathbf{u}_3
\]

Thus in terms of the basis \( \gamma \equiv \{ \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3 \} \), the matrix of this transformation is
\[
[T]_\gamma \equiv \begin{pmatrix}
cos \theta & -\sin \theta & 0 \\
\sin \theta & \cos \theta & 0 \\
0 & 0 & 1
\end{pmatrix}.
\]

This is not desirable because it involves a funny basis. I want to obtain the matrix of the transformation in terms of the usual basis \( \beta \equiv \{ \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3 \} \) because it is in terms of this basis that we usually deal with vectors in \( \mathbb{R}^3 \). From Proposition \ref{prop:rot}, if \( [T]_\beta \) is this matrix,
\[
= \begin{pmatrix}
\mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3
\end{pmatrix}^{-1} [T]_\beta \begin{pmatrix}
\mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3
\end{pmatrix}
\]

and so you can solve for \( [T]_\beta \) if you know the \( \mathbf{u}_i \).

Recall why this is so.

\[
\mathbb{R}^3 \xrightarrow{[T]_\gamma} \mathbb{R}^3 \\
\mathbb{R}^3 \xrightarrow{T} \mathbb{R}^3 \\
\mathbb{R}^3 \xrightarrow{[T]_\beta} \mathbb{R}^3
\]

The map \( q_\gamma \) is accomplished by a multiplication on the left by \( \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix} \). Thus
\[
[T]_\beta = q_\gamma [T]_\gamma q_\gamma^{-1} = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix} [T]_\gamma \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix}^{-1}.
\]

Suppose the unit vector \( \mathbf{u}_3 \) about which the counterclockwise rotation takes place is \( (\alpha, \beta, \gamma) \). Then I obtain vectors, \( \mathbf{u}_1 \) and \( \mathbf{u}_2 \) such that \( \{ \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3 \} \) is a right handed orthonormal system with \( \mathbf{u}_3 = (\alpha, \beta, \gamma) \) and then use the above result. It is of course somewhat arbitrary how this is
accomplished. I will assume however, that $|c| \neq 1$ since otherwise you are looking at either clockwise or counter clockwise rotation about the positive $z$ axis and this is a problem which is fairly easy. Indeed, the matrix of such a rotation in terms of the usual basis is just

\[
\begin{pmatrix}
\cos \theta & -\sin \theta & 0 \\
\sin \theta & \cos \theta & 0 \\
0 & 0 & 1
\end{pmatrix}
\] (5.6)

Then let $\mathbf{u}_3 = (a, b, c)$ and $\mathbf{u}_2 \equiv \frac{1}{\sqrt{a^2 + b^2}} (b, -a, 0)$. This one is perpendicular to $\mathbf{u}_3$. If $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is to be a right hand system it is necessary to have

\[
\mathbf{u}_1 = \mathbf{u}_2 \times \mathbf{u}_3 = \frac{1}{\sqrt{(a^2 + b^2)(a^2 + b^2 + c^2)}} (-ac, -bc, a^2 + b^2)
\]

Now recall that $\mathbf{u}_3$ is a unit vector and so the above equals

\[
\frac{1}{\sqrt{(a^2 + b^2)}} (-ac, -bc, a^2 + b^2)
\]

Then from the above, $A$ is given by

\[
\begin{pmatrix}
\frac{-ac}{\sqrt{(a^2 + b^2)}} & \frac{b}{\sqrt{a^2 + b^2}} & a \\
\frac{-bc}{\sqrt{(a^2 + b^2)}} & \frac{-a}{\sqrt{a^2 + b^2}} & b \\
\frac{0}{\sqrt{a^2 + b^2}} & c & 0
\end{pmatrix}
\begin{pmatrix}
\cos \theta & -\sin \theta & 0 \\
\sin \theta & \cos \theta & 0 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
\frac{-ac}{\sqrt{(a^2 + b^2)}} & \frac{b}{\sqrt{a^2 + b^2}} & a \\
\frac{-bc}{\sqrt{(a^2 + b^2)}} & \frac{-a}{\sqrt{a^2 + b^2}} & b \\
\frac{0}{\sqrt{a^2 + b^2}} & c & 0
\end{pmatrix}^{-1}
\]

It is easy to take the inverse of this matrix on the left. You can check right away that its inverse is nothing but its transpose. Then doing the computation and then some simplification yields

\[
= \begin{pmatrix}
a^2 + (1 - a^2) \cos \theta & ab (1 - \cos \theta) - c \sin \theta & ac (1 - \cos \theta) + b \sin \theta \\
ab (1 - \cos \theta) + c \sin \theta & b^2 + (1 - b^2) \cos \theta & bc (1 - \cos \theta) - a \sin \theta \\
ac (1 - \cos \theta) - b \sin \theta & bc (1 - \cos \theta) + a \sin \theta & c^2 + (1 - c^2) \cos \theta
\end{pmatrix}^{-1}
\] (5.7)

With this, it is clear how to rotate clockwise about the unit vector, $(a, b, c)$. Just rotate counter clockwise through an angle of $-\theta$. Thus the matrix for this clockwise rotation is just

\[
= \begin{pmatrix}
a^2 + (1 - a^2) \cos \theta & ab (1 - \cos \theta) + c \sin \theta & ac (1 - \cos \theta) - b \sin \theta \\
ab (1 - \cos \theta) - c \sin \theta & b^2 + (1 - b^2) \cos \theta & bc (1 - \cos \theta) + a \sin \theta \\
ac (1 - \cos \theta) + b \sin \theta & bc (1 - \cos \theta) - a \sin \theta & c^2 + (1 - c^2) \cos \theta
\end{pmatrix}
\]

In deriving it was assumed that $c \neq \pm 1$ but even in this case, it gives the correct answer. Suppose for example that $c = 1$ so you are rotating in the counterclockwise direction about the positive $z$ axis. Then $a, b$ are both equal to zero and reduces to $\Box$.

### 5.4 Exercises

1. If $A, B,$ and $C$ are each $n \times n$ matrices and $ABC$ is invertible, why are each of $A, B,$ and $C$ invertible?

2. Give an example of a $3 \times 2$ matrix with the property that the linear transformation determined by this matrix is one to one but not onto.

3. Explain why $Ax = 0$ always has a solution whenever $A$ is a linear transformation.

4. Recall that a line in $\mathbb{R}^n$ is of the form $x + tv$ where $t \in \mathbb{R}$. Recall that $v$ is a “direction vector”. Show that if $T : \mathbb{R}^n \to \mathbb{R}^m$ is linear, then the image of $T$ is either a line or a point.
5. In the following examples, a linear transformation, $T$ is given by specifying its action on a basis $\beta$. Find its matrix with respect to this basis.

(a) $T \begin{pmatrix} 1 \\ 2 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 2 \end{pmatrix} + 1 \begin{pmatrix} -1 \\ 1 \end{pmatrix}, T \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$

(b) $T \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} + 1 \begin{pmatrix} -1 \\ 1 \end{pmatrix}, T \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

(c) $T \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 2 \end{pmatrix} + 1 \begin{pmatrix} 1 \\ 0 \end{pmatrix}, T \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 2 \end{pmatrix}$

6. In each example above, find a matrix $A$ such that for every $x \in \mathbb{R}^2$, $Tx = Ax$.

7. Consider the linear transformation $T_\theta$ which rotates every vector in $\mathbb{R}^2$ through the angle of $\theta$. Find the matrix $A_\theta$ such that $T_\theta x = A_\theta x$. **Hint:** You need to have the columns of $A_\theta$ be $Te_1$ and $Te_2$. Review why this is before using this. Then simply find these vectors from trigonometry.

8. If you did the above problem right, you got

$$A_\theta = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

Derive the famous trig. identities for the sum of two angles by using the fact that $A_{\theta + \phi} = A_\theta A_\phi$ and the above description.

9. Let $\beta = \{u_1, \cdots, u_n\}$ be a basis for $\mathbb{F}^n$ and let $T: \mathbb{F}^n \to \mathbb{F}^n$ be defined as follows.

$$T \left( \sum_{k=1}^{n} a_k u_k \right) = \sum_{k=1}^{n} a_k b_k u_k$$

First show that $T$ is a linear transformation. Next show that the matrix of $T$ with respect to this basis is $[T]_\beta = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$

Show that the above definition is equivalent to simply specifying $T$ on the basis vectors of $\beta$ by

$$T(u_k) = b_k u_k.$$ 

10. Let $T$ be given by specifying its action on the vectors of a basis $\beta = \{u_1, \cdots, u_n\}$ as follows.

$$T u_k = \sum_{j=1}^{n} a_{jk} u_j.$$ 

Letting $A = (a_{ij})$, verify that $[T]_\beta = A$. It is done in the chapter, but go over it yourself. 

Show that $[T]_\gamma = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix} [T]_\beta \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix}^{-1}$ (5.8)

11. Let $a$ be a fixed vector. The function $T_a$ defined by $T_a v = a + v$ has the effect of translating all vectors by adding $a$. Show this is not a linear transformation. Explain why it is not possible to realize $T_a$ in $\mathbb{R}^3$ by multiplying by a $3 \times 3$ matrix.
12. In spite of Problem 11, we can represent both translations and rotations by matrix multiplication at the expense of using higher dimensions. This is done by the homogeneous coordinates. I will illustrate in $\mathbb{R}^3$ where most interest in this is found. For each vector $v = (v_1, v_2, v_3)^T$, consider the vector in $\mathbb{R}^4 (v_1, v_2, v_3, 1)^T$. What happens when you do

$$
\begin{pmatrix}
1 & 0 & 0 & a_1 \\
0 & 1 & 0 & a_2 \\
0 & 0 & 1 & a_3 \\
0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
v_1 \\
v_2 \\
v_3 \\
1
\end{pmatrix}
$$

Describe how to consider both rotations and translations all at once by forming appropriate $4 \times 4$ matrices.

13. You want to add $\left( 1, \ 2, \ 3 \right)$ to every point in $\mathbb{R}^3$ and then rotate about the $z$ axis counterclockwise through an angle of $30^\circ$. Find what happens to the point $\left( 1, \ 1, \ 1 \right)$.

14. Let $P_3$ denote the set of real polynomials of degree no more than 3, defined on an interval $[a, b]$. Show that $P_3$ is a subspace of the vector space of all functions defined on this interval. Show that a basis for $P_3$ is $\{1, x, x^2, x^3\}$. Now let $D$ denote the differentiation operator which sends a function to its derivative. Show $D$ is a linear transformation which sends $P_3$ to $P_3$. Find the matrix of this linear transformation with respect to the given basis.

15. Generalize the above problem to $P_n$, the space of polynomials of degree no more than $n$ with basis $\{1, x, \cdots, x^n\}$.

16. If $A$ is an $n \times n$ invertible matrix, show that $A^T$ is also and that in fact, $(A^T)^{-1} = (A^{-1})^T$.

17. Suppose you have an invertible $n \times n$ matrix $A$. Consider the polynomials

$$
\begin{pmatrix}
p_1 (x) \\
\vdots \\
p_n (x)
\end{pmatrix}
= A
\begin{pmatrix}
1 \\
\vdots \\
x^{n-1}
\end{pmatrix}
$$

Show that these polynomials $p_1 (x), \cdots, p_n (x)$ are a linearly independent set of functions.

18. Let the linear transformation be $T = D^2 + 1$, defined as $T f = f'' + f$. Find the matrix of this linear transformation with respect to the given basis $\{1, x, x^2, x^3\}$.

19. Let $L$ be the linear transformation taking polynomials of degree at most three to polynomials of degree at most three given by

$$
D^2 + 2D + 1
$$

where $D$ is the differentiation operator. Find the matrix of this linear transformation relative to the basis $\{1, x, x^2, x^3\}$. Find the matrix directly and then find the matrix with respect to the differential operator $D + 1$ and multiply this matrix by itself. You should get the same thing. Why?

20. Let $L$ be the linear transformation taking polynomials of degree at most three to polynomials of degree at most three given by $D^2 + 5D + 4$ where $D$ is the differentiation operator. Find the matrix of this linear transformation relative to the bases $\{1, x, x^2, x^3\}$. Find the matrix directly and then find the matrices with respect to the differential operators $D + 1, D + 4$ and multiply these two matrices. You should get the same thing. Why?

21. Suppose $A \in \mathcal{L}(V, W)$ where $\dim(V) > \dim(W)$. Show $\ker(A) \neq \{0\}$. That is, show there exist nonzero vectors $v \in V$ such that $Av = 0$. 

22. A vector \( v \) is in the convex hull of a nonempty set if there are finitely many vectors of \( S, \{v_1, \ldots, v_m\} \) and nonnegative scalars \( \{t_1, \ldots, t_m\} \) such that

\[
v = \sum_{k=1}^{m} t_k v_k, \quad \sum_{k=1}^{m} t_k = 1.
\]

Such a linear combination is called a convex combination. Suppose now that \( S \subseteq V \), a vector space of dimension \( n \). Show that if \( v = \sum_{k=1}^{m} t_k v_k \) is a vector in the convex hull for \( m > n + 1 \), then there exist other scalars \( \{t'_k\} \) such that

\[
v = \sum_{k=1}^{m-1} t'_k v_k.
\]

Thus every vector in the convex hull of \( S \) can be obtained as a convex combination of at most \( n + 1 \) points of \( S \). This incredible result is in Rudin [29]. **Hint:** Consider \( L : \mathbb{R}^m \to V \times \mathbb{R} \) defined by

\[
L(a) = \left( \sum_{k=1}^{m} a_k v_k, \sum_{k=1}^{m} a_k \right)
\]

Explain why \( \ker(L) \neq \{0\} \). Next, letting \( a \in \ker(L) \setminus \{0\} \) and \( \lambda \in \mathbb{R} \), note that \( \lambda a \in \ker(L) \). Thus for all \( \lambda \in \mathbb{R} \),

\[
v = \sum_{k=1}^{m} (t_k + \lambda a_k) v_k.
\]

Now vary \( \lambda \) till some \( t_k + \lambda a_k = 0 \) for some \( a_k \neq 0 \).

23. For those who know about compactness, use Problem 22 to show that if \( S \subseteq \mathbb{R}^n \) and \( S \) is compact, then so is its convex hull.

24. Show that if \( L \in \mathcal{L}(V, W) \) (linear transformation) where \( V \) and \( W \) are vector spaces, then if \( L y_p = f \) for some \( y_p \in V \), then the general solution of \( Ly = f \) is of the form \( \ker(L) + y_p \).

25. Suppose \( Ax = b \) has a solution. Explain why the solution is unique precisely when \( Ax = 0 \) has only the trivial (zero) solution.

26. Let \( L : \mathbb{R}^n \to \mathbb{R} \) be linear. Show that there exists a vector \( a \in \mathbb{R}^n \) such that \( Ly = a^T y \).

27. Let the linear transformation \( T \) be determined by

\[
T(x) = \begin{pmatrix} 1 & 0 & -5 & -7 \\ 0 & 1 & -3 & -9 \\ 1 & 1 & -8 & -16 \end{pmatrix} x
\]

Find the rank of this transformation.

28. Let \( Tf = (D^2 + 5D + 4) f \) for \( f \) in the vector space of polynomials of degree no more than 3 where we consider \( T \) to map into the same vector space. Find the rank of \( T \). You might want to use Proposition 1.3.1.

29. (Extra important) Let \( A \) be an \( n \times n \) matrix. The trace of \( A \), \( \text{trace}(A) \) is defined as \( \sum_i A_{ii} \). It is just the sum of the entries on the main diagonal. Show \( \text{trace}(A) = \text{trace}(A^T) \). Suppose \( A \) is \( m \times n \) and \( B \) is \( n \times m \). Show that \( \text{trace}(AB) = \text{trace}(BA) \). Now show that if \( A \) and \( B \) are similar \( n \times n \) matrices, then \( \text{trace}(A) = \text{trace}(B) \). Recall that \( A \) is similar to \( B \) means \( A = S^{-1}BS \) for some matrix \( S \).
30. Suppose you have a monic polynomial \( \phi (\lambda) \) which is irreducible over \( \mathbb{F} \) the field of scalars. Remember that this means that its greatest common divisor is 1. Equivalently, the only polynomials which divide \( \phi (\lambda) \) are multiples of \( \phi (\lambda) \) and scalars. Say

\[
\phi (\lambda) = a_0 + a_1 \lambda + \cdots + a_{d-1} \lambda^{d-1} + \lambda^d
\]

Now consider \( A \in \mathcal{L}(V,V) \) where \( V \) is a vector space. Consider \( \ker (\phi (A)) \) and suppose this is not 0. For \( x \in \ker (\phi (A)), x \neq 0 \), let \( \beta_x = \{ x, Ax, \cdots, A^{d-1}x \} \). Show that \( \beta_x \) is an independent set of vectors if \( x \neq 0 \).

31. Let \( V \) be a finite dimensional vector space and let \( A \in \mathcal{L}(V,V) \). Also let \( W \) be a subspace of \( V \) such that \( A(W) \subseteq W \). We call such a subspace an \( A \) invariant subspace. Say \( \{ w_1, \cdots, w_s \} \) is a basis for \( W \). Also let \( x \in U \cap W \) where \( U \) is an \( A \) invariant subspace which is contained in \( \ker (\phi (A)) \). Then you know that \( \{ w_1, \cdots, w_s, x \} \) is linearly independent. Show that in fact \( \{ w_1, \cdots, w_s, \beta_x \} \) is linearly independent where \( \beta_x \) is given in the above problem. Hint: Suppose you have

\[
\sum_{k=1}^s a_kw_k + \sum_{j=1}^d b_jA^{j-1}x = 0. \tag{*}
\]

You need to verify that the second sum is 0. From this it will follow that each \( b_j \) is 0 and then each \( a_k = 0 \). Let \( S = \sum_{j=1}^d b_jA^{j-1}x \). Observe that \( \beta_S \subseteq \beta_x \) and if \( S \neq 0 \), then \( \beta_S \) is independent from the above problem and both \( \beta_x \) and \( \beta_S \) have the same dimension. You will argue that \( \text{span}(\beta_S) \subseteq W \cap \text{span}(\beta_x) \subseteq \text{span}(\beta_x) \) and then use Problem 2 on Page 72.

32. In the situation of the above problem, show that there exist finitely many vectors in \( U \), \( \{ x_1, \cdots, x_m \} \) such that \( \{ w_1, \cdots, w_s, \beta_{x_1}, \cdots, \beta_{x_m} \} \) is a basis for \( U + W \). The last vector space is defined as the set of all \( y + w \) where \( y \in U \) and \( w \in W \).

33. In the situation of the above where \( \phi (\lambda) \) is irreducible. Let \( U \) be defined as

\[ U = \phi (A) (\ker (\phi (A)^m) ) \]

Explain why \( U \subseteq \ker (\phi (A)^{m-1}) \). Suppose you have a linearly independent set in \( U \) which is \( \{ \beta_{x_1}, \cdots, \beta_{x_r} \} \). Here the notation means

\[ \beta_x = \{ x, Ax, \cdots, A^{m-1}x \} \]

where these vectors are independent but \( A^m x \) is in the span of these. Such exists any time you have \( x \in \ker (g(A)) \) for \( g(\lambda) \) a polynomial. Letting \( \phi (A)y_i = x_i \), explain why \( \{ \beta_{y_1}, \cdots, \beta_{y_r} \} \) is also linearly independent. This is like the theorem presented earlier that the inverse image of a linearly independent set is linearly independent but it is more complicated here because instead of single vectors, we are considering sets \( \beta_x \).
Chapter 6

Direct Sums And Block Diagonal Matrices

This is a convenient place to put a very interesting result about direct sums and block diagonal matrices. First is the notion of a direct sum. In all of this, $V$ will be a finite dimensional vector space of dimension $n$ and field of scalars $\mathbb{F}$.

**Definition 6.0.1** Let $\{V_i\}_{i=1}^r$ be subspaces of $V$. Then

$$
\sum_{i=1}^r V_i \equiv V_1 + \cdots + V_r
$$

denotes all sums of the form $\sum_{i=1}^r v_i$ where $v_i \in V_i$. If whenever

$$
\sum_{i=1}^r v_i = 0, v_i \in V_i, \quad (6.1)
$$

it follows that $v_i = 0$ for each $i$, then a special notation is used to denote $\sum_{i=1}^r V_i$. This notation is

$$
V_1 \oplus \cdots \oplus V_r,
$$

and it is called a direct sum of subspaces. A subspace $W$ of $V$ is called $A$ invariant for $A \in \mathcal{L}(V,V)$ if $AW \subseteq W$.

The important idea is that you seek to understand $A$ by looking at what it does on each $V_i$. It is a lot like knowing $A$ by knowing what it does to a basis, an idea used earlier.

**Lemma 6.0.2** If $V = V_1 \oplus \cdots \oplus V_r$ and if $\beta_i = \{v_{i1}^1, \cdots, v_{im_i}^i\}$ is a basis for $V_i$, then a basis for $V$ is $\{\beta_1, \cdots, \beta_r\}$. Thus

$$
\dim (V) = \sum_{i=1}^r \dim (V_i).
$$

Conversely, if $\beta_i$ linearly independent and if a basis for $V$ is $\{\beta_1, \cdots, \beta_r\}$, then

$$
V = \text{span} (\beta_1) \oplus \cdots \oplus \text{span} (\beta_r)
$$

**Proof:** Suppose $\sum_{i=1}^r \sum_{j=1}^{m_i} c_{ij} v_j^i = 0$. then since it is a direct sum, it follows for each $i$,

$$
\sum_{j=1}^{m_i} c_{ij} v_j^i = 0
$$

117
and now since \( \{v^1_i, \ldots, v^m_i\} \) is a basis, each \( c_{ij} = 0 \) for each \( j \), this for each \( i \).

Suppose now that each \( \beta_i \) is independent and a basis is \( \{\beta_1, \ldots, \beta_r\} \). Then clearly
\[
V = \text{span} (\beta_1) + \cdots + \text{span} (\beta_r)
\]
Suppose then that \( 0 = \sum_{i=1}^r \sum_{j=1}^{m_i} c_{ij} v^i_j \), the inside sum being something in \( \text{span} (\beta_i) \). Since \( \{\beta_1, \ldots, \beta_r\} \) is a basis, each \( c_{ij} = 0 \). Thus each \( \sum_{j=1}^{m_i} c_{ij} v^i_j = 0 \) and so \( V = \text{span} (\beta_1) + \cdots + \text{span} (\beta_r) \).

Thus, from this lemma, we can produce a basis for \( V \) of the form \( \{\beta_1, \ldots, \beta_r\} \), so what is the matrix of a linear transformation \( A \) such that each \( V_i \) is \( A \) invariant?

**Theorem 6.0.3** Suppose \( V \) is a vector space with field of scalars \( \mathbb{F} \) and \( A \in \mathcal{L}(V, V) \). Suppose also
\[
V = V_1 \oplus \cdots \oplus V_q
\]
where each \( V_k \) is \( A \) invariant. \( (AV_k \subseteq V_k) \) Also let \( \beta_k \) be an ordered basis for \( V_k \) and let \( A_k \) denote the restriction of \( A \) to \( V_k \). Letting \( M^k \) denote the matrix of \( A_k \) with respect to this basis, it follows the matrix of \( A \) with respect to the basis \( \{\beta_1, \ldots, \beta_q\} \) is
\[
\begin{pmatrix}
M^1 & 0 & \cdots & 0 \\
0 & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots \\
0 & \cdots & 0 & M^q
\end{pmatrix}
\]

**Proof:** Let \( \beta \) denote the ordered basis \( \{\beta_1, \ldots, \beta_q\} \), \( |\beta_k| \) being the number of vectors in \( \beta_k \). Let \( q_k : \mathbb{F}^{|\beta_k|} \to V_k \) be the usual map such that the following diagram commutes.

\[
\begin{array}{ccc}
A_k & \to & V_k \\
q_k & \uparrow & \uparrow q_k \\
\mathbb{F}^{|\beta_k|} & \to & \mathbb{F}^{|\beta_k|}
\end{array}
\]

Thus \( A_k q_k = q_k M^k \). Then if \( q \) is the map from \( \mathbb{F}^n \) to \( V \) corresponding to the ordered basis \( \beta \) just described,
\[
q \begin{pmatrix} 0 & \cdots & x & \cdots & 0 \end{pmatrix}^T = q_k x,
\]
where \( x \) occupies the positions between \( \sum_{i=1}^{k-1} |\beta_i| + 1 \) and \( \sum_{i=1}^k |\beta_i| \). Then \( M \) will be the matrix of \( A \) with respect to \( \beta \) if and only if a similar diagram to the above commutes. Thus it is required that \( Aq = qM \). However, from the description of \( q \) just made, and the invariance of each \( V_k \),
\[
A q \begin{pmatrix} 0 \\
\vdots \\
x \\
\vdots \\
0 \end{pmatrix} = A_k q_k x = q_k M^k x = q \begin{pmatrix} M^1 \\
\ddots \\
M^k \\
\ddots \\
0 \end{pmatrix} \begin{pmatrix} 0 \\
\vdots \\
x \\
\vdots \\
0 \end{pmatrix}
\]
It follows that the above block diagonal matrix is the matrix of \( A \) with respect to the given ordered basis.

The matrix of \( A \) with respect to the ordered basis \( \beta \) which is described above is called a block diagonal matrix. Sometimes the blocks consist of a single number.
Example 6.0.4  Consider the following matrix.

\[ A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & -1 \\ -2 & 2 & 3 \end{pmatrix} \]

Let \( V_1 \equiv \text{span} \left( \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right) \), \( V_2 \equiv \text{span} \left( \begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix} \right) \). Show that \( \mathbb{R}^3 = V_1 \oplus V_2 \) and that \( V_i \) is \( A \)-invariant. Find the matrix of \( A \) with respect to the ordered basis \( \left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \right\} \) (**).

First note that

\[
\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & -1 \\ -2 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}
\]

Therefore, \( A(V_1) \subseteq V_1 \). Similarly,

\[
\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & -1 \\ -2 & 2 & 3 \end{pmatrix} \begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix} = 2 \begin{pmatrix} 0 \\ -2 \\ 4 \end{pmatrix} = 2 \begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix}
\]

and so \( A(V_2) \subseteq V_2 \). The vectors in (***) clearly are a basis for \( \mathbb{R}^3 \). You can verify this by observing that there is a unique solution \( x, y, z \) to the system of equations

\[
x \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + y \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + z \begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}
\]

for any choice of the right side. Therefore, by Lemma 6.0.2, \( \mathbb{R}^3 = V_1 \oplus V_2 \).

If you look at the restriction of \( A \) to \( V_1 \), what is the matrix of this restriction? It satisfies

\[
\begin{pmatrix} A \left( \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right) \right) = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \right) = \begin{pmatrix} a & b \\ c & d \end{pmatrix}
\]

Thus, from what was observed above, you need the matrix on the right to satisfy

\[
\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \right) = \begin{pmatrix} a & b \\ c & d \end{pmatrix}
\]

and so the matrix on the right is just \( \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \). As to the matrix of \( A \) restricted to \( V_2 \), we need

\[
A \begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix} = a \begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix}
\]
where $a$ is a $1 \times 1$ matrix. Thus $a = 2$ and so the matrix of $A$ with respect to the ordered basis given above is

$$
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 2
\end{pmatrix}
$$

What if you changed the order of the vectors in the basis? Suppose you had them ordered as

$$
\begin{pmatrix}
1 \\
1 \\
0
\end{pmatrix}, \begin{pmatrix}
0 \\
-1 \\
2
\end{pmatrix}, \begin{pmatrix}
1 \\
0 \\
1
\end{pmatrix}
$$

Then you would have three invariant subspaces whose direct sum is $\mathbb{R}^3$,

$$
\text{span} \begin{pmatrix}
1 \\
1 \\
0
\end{pmatrix}, \text{span} \begin{pmatrix}
0 \\
-1 \\
2
\end{pmatrix}, \text{and } \text{span} \begin{pmatrix}
1 \\
0 \\
1
\end{pmatrix}
$$

Then the matrix of $A$ with respect to this ordered basis is

$$
\begin{pmatrix}
1 & 0 & 0 \\
0 & 2 & 0 \\
0 & 0 & 1
\end{pmatrix}
$$

**Example 6.0.5** Consider the following matrix.

$$
A = \begin{pmatrix}
3 & 1 & 0 \\
-1 & 1 & 0 \\
-1 & -1 & 1
\end{pmatrix}
$$

Let

$$
V_1 \equiv \text{span} \begin{pmatrix}
0 \\
0 \\
1
\end{pmatrix}, V_2 \equiv \text{span} \begin{pmatrix}
1 \\
0 \\
-1
\end{pmatrix}
$$

Show that these are $A$ invariant subspaces and find the matrix of $A$ with respect to the ordered basis

$$
\begin{pmatrix}
0 \\
0 \\
1
\end{pmatrix}, \begin{pmatrix}
1 \\
-1 \\
0
\end{pmatrix}, \begin{pmatrix}
1 \\
0 \\
-1
\end{pmatrix}
$$

First note that

$$
\begin{pmatrix}
3 & 1 & 0 \\
-1 & 1 & 0 \\
-1 & -1 & 1
\end{pmatrix} - 2 \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix} \begin{pmatrix}
1 \\
0 \\
-1
\end{pmatrix} = \begin{pmatrix}
1 \\
-1 \\
0
\end{pmatrix}
$$

and so $A \begin{pmatrix}
1 \\
0 \\
-1
\end{pmatrix}$ is in the span of

$$
\begin{pmatrix}
1 \\
-1 \\
0
\end{pmatrix}, \begin{pmatrix}
1 \\
0 \\
-1
\end{pmatrix}
$$
Also
\[
\begin{pmatrix}
3 & 1 & 0 \\
-1 & 1 & 0 \\
-1 & -1 & 1
\end{pmatrix}
\begin{pmatrix}
1 \\
-1 \\
0
\end{pmatrix}
= \begin{pmatrix}
2 \\
-2 \\
0
\end{pmatrix}
\in \text{span}\left(\begin{pmatrix}
1 \\
-1 \\
0
\end{pmatrix}, \begin{pmatrix}
1 \\
0 \\
-1
\end{pmatrix}\right)
\]
Thus \(V_2\) is \(A\) invariant. What is the matrix of \(A\) restricted to \(V_2\)? We need
\[
\begin{pmatrix}
A \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, A \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}\end{pmatrix} = \begin{pmatrix}
1 \\
-1 \\
0
\end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix}
\]
Now it was shown above that
\[
A \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} + \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}
\]
and so the matrix is of the form
\[
\begin{pmatrix}
a & 1 \\
c & 2
\end{pmatrix}
\]
Then it was also shown that \(A \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}\) and so the matrix is of the form
\[
\begin{pmatrix}
2 & 1 \\
0 & 2
\end{pmatrix}
\]
As to \(V_1\),
\[
A \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}
\]
and the matrix of \(A\) restricted to \(V_1\) is just the \(1 \times 1\) matrix consisting of the number 1. Thus the matrix of \(A\) with respect to this basis is
\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 2 & 1 \\
0 & 0 & 2
\end{pmatrix}
\]
How can you come up with invariant subspaces? In the next section, I will give a systematic way based on a profound theorem of Sylvester. However, there is also a very easy way to come up with an invariant subspace. Let \(v \in V\) an \(n\) dimensional vector space and let \(A \in \mathcal{L}(V, V)\). Let \(W \equiv \text{span}(v, Av, A^2v, \ldots)\). It is left as an exercise to verify that \(W\) is a finite dimensional subspace of \(V\). Recall that the span is the set of all finite linear combinations. Of course \(W\) might be all of \(V\) or it might be a proper subset of \(V\). The method of Sylvester will end up typically giving proper invariant subspaces whose direct sum is the whole space.
6.1 A Theorem Of Sylvester, Direct Sums

The notation is defined as follows. First recall the definition of \( \ker \) in Problem 23 on Page 62.

**Definition 6.1.1** Let \( L \in \mathcal{L}(V, W) \). Then \( \ker(L) = \{ v \in V : Lv = 0 \} \).

**Lemma 6.1.2** Whenever \( L \in \mathcal{L}(V, W) \), \( \ker(L) \) is a subspace. Also, if \( V \) is an \( n \) dimensional vector space and \( W \) is a subspace of \( V \), then \( W = V \) if and only if \( \dim(W) = n \).

**Proof:** If \( a, b \) are scalars and \( v, w \) are in \( \ker(L) \), then

\[
L(av + bw) = aL(v) + bL(w) = 0 + 0 = 0
\]

As to the last claim, it is clear that \( \dim(W) \leq n \). If \( \dim(W) = n \), then, letting \( \{w_1, \ldots, w_n\} \) be a basis for \( W \), there can be no \( v \in V \setminus W \) because then \( v \notin \text{span}(w_1, \ldots, w_n) \) and so by Lemma 6.1.2 \( \{w_1, \ldots, w_n, v\} \) would be independent which is impossible by Theorem 3.1.5. You have an independent set which is longer than a spanning set.

Suppose now that \( A \in \mathcal{L}(V, W) \) and \( B \in \mathcal{L}(W, U) \) where \( V, W, U \) are all finite dimensional vector spaces. Then it is interesting to consider \( \ker(BA) \). The following theorem of Sylvester is a very useful and important result.

**Theorem 6.1.3** Let \( A \in \mathcal{L}(V, W) \) and \( B \in \mathcal{L}(W, U) \) where \( V, W, U \) are all vector spaces over a field \( F \). Suppose also that \( \ker(A) \) and \( A(\ker(BA)) \) are finite dimensional subspaces. Then

\[
\dim(\ker(BA)) \leq \dim(\ker(B)) + \dim(\ker(A))
\]

Equality holds if and only if \( A(\ker(BA)) = \ker(B) \).

**Proof:** If \( x \in \ker(BA) \), then \( Ax \in \ker(B) \) and so

\[
A(\ker(BA)) \subseteq \ker(B).
\]

The following picture may help.

Now let \( \{x_1, \ldots, x_n\} \) be a basis of \( \ker(A) \) and let \( \{Ay_1, \ldots, Ay_m\} \) be a basis for \( A(\ker(BA)) \). Take any \( z \in \ker(BA) \). Then \( Az = \sum_{i=1}^{m} a_iAy_i \) and so

\[
A\left(z - \sum_{i=1}^{m} a_iy_i\right) = 0
\]

which means \( z - \sum_{i=1}^{m} a_iy_i \in \ker(A) \) and so there are scalars \( b_i \) such that

\[
z - \sum_{i=1}^{m} a_iy_i = \sum_{j=1}^{n} b_ix_i.
\]

It follows \( \text{span}(x_1, \ldots, x_n, y_1, \ldots, y_m) \supseteq \ker(BA) \) and so by the first part, (See the picture.)

\[
\dim(\ker(BA)) \leq n + m \leq \dim(\ker(A)) + \dim(\ker(B))
\]

Now \( \{x_1, \ldots, x_n, y_1, \ldots, y_m\} \) is linearly independent because if

\[
\sum_i a_ix_i + \sum_j b_jy_j = 0
\]
then you could do $A$ to both sides and conclude that $\sum_{j} b_j A y_j = 0$ which requires that each $b_j = 0$. Then it follows that each $a_i = 0$ also because it implies $\sum_{i} a_i x_i = 0$. Thus

$$\{x_1, \cdots, x_n, y_1, \cdots, y_m\}$$

is a basis for $\ker(BA)$. Then by Lemma 6.1.4 $A(\ker(BA)) = \ker(B)$ if and only if $m = \dim(\ker(B))$ if and only if

$$\dim(\ker(BA)) = m + n = \dim(\ker(B)) + \dim(\ker(A)).$$

Of course this result holds for any finite product of linear transformations by induction. One way this is quite useful is in the case where you have a finite product of linear transformations $\prod_{i=1}^{p} L_i$ all in $\mathcal{L}(V;V)$. Then

$$\dim\left(\ker\prod_{i=1}^{l} L_i\right) \leq \sum_{i=1}^{l} \dim(\ker L_i).$$

Now here is a useful lemma which is likely already understood.

**Lemma 6.1.4** Let $L \in \mathcal{L}(V,W)$ where $V,W$ are $n$ dimensional vector spaces. Then $L$ is one to one, if and only if $L$ is also onto. In fact, if $\{v_1, \cdots, v_n\}$ is a basis, then so is $\{Lv_1, \cdots, Lv_n\}$.

**Proof:** Let $\{v_1, \cdots, v_n\}$ be a basis for $V$. Then I claim that $\{Lv_1, \cdots, Lv_n\}$ is a basis for $W$. First of all, I show $\{Lv_1, \cdots, Lv_n\}$ is linearly independent. Suppose

$$\sum_{k=1}^{n} c_kLv_k = 0.$$

Then

$$L\left(\sum_{k=1}^{n} c_kv_k\right) = 0$$

and since $L$ is one to one, it follows

$$\sum_{k=1}^{n} c_kv_k = 0$$

which implies each $c_k = 0$. Therefore, $\{Lv_1, \cdots, Lv_n\}$ is linearly independent. If there exists $w$ not in the span of these vectors, then by Lemma 6.1.4 $\{Lv_1, \cdots, Lv_n, w\}$ would be independent and this contradicts the exchange theorem, Theorem 3.1.7 because it would be a linearly independent set having more vectors than the spanning set $\{v_1, \cdots, v_n\}$.

Conversely, suppose $L$ is onto. Then there exists a basis for $W$ which is of the form $\{Lv_1, \cdots, Lv_n\}$. It follows that $\{v_1, \cdots, v_n\}$ is linearly independent. Hence it is a basis for $V$ by similar reasoning to the above. Then if $Lx = 0$, it follows that there are scalars $c_i$ such that $x = \sum c_i v_i$ and consequently $0 = Lx = \sum c_i Lv_i$. Therefore, each $c_i = 0$ and so $x = 0$ also. Thus $L$ is one to one. ■

Here is a fundamental lemma which gives a typical situation where a vector space is the direct sum of subspaces.

**Lemma 6.1.5** Let $L_i$ be in $\mathcal{L}(V,V)$ and suppose for $i \neq j$, $L_iL_j = L_jL_i$ and also $L_i$ is one to one on $\ker(L_j)$ whenever $i \neq j$. Then

$$\ker\left(\prod_{i=1}^{p} L_i\right) = \ker(L_1) + \cdots + \ker(L_p)$$

Here $\prod_{i=1}^{p} L_i$ is the product of all the linear transformations. It signifies

$$L_p \circ L_{p-1} \circ \cdots \circ L_1$$

or the product in any other order since the transformations commute.
**Proof:** Note that since the operators commute, \( L_j : \ker(L_i) \rightarrow \ker(L_i) \). Here is why. If \( L_iy = 0 \) so that \( y \in \ker(L_i) \), then

\[
L_iL_jy = L_jL_iy = L_j0 = 0
\]

and so \( L_j : \ker(L_i) \mapsto \ker(L_i) \). Next observe that it is obvious that, since the operators commute,

\[
\sum_{i=1}^{p} \ker(L_i) \subseteq \ker\left(\prod_{i=1}^{p} L_i\right)
\]

Next, why is \( \sum_{i=1}^{p} \ker(L_i) = \ker(L_1) \oplus \cdots \oplus \ker(L_p) \)? Suppose

\[
\sum_{i=1}^{p} v_i = 0, \ v_i \in \ker(L_i),
\]

but some \( v_i \neq 0 \). Then do \( \prod_{j \neq i} L_j \) to both sides. Since the linear transformations commute, this results in

\[
\left(\prod_{j \neq i} L_j\right)\left( v_i \right) = 0
\]

which contradicts the assumption that these \( L_j \) are one to one on \( \ker(L_i) \) and the observation that they map \( \ker(L_i) \) to \( \ker(L_i) \). Thus if

\[
\sum_{i} v_i = 0, \ v_i \in \ker(L_i)
\]

then each \( v_i = 0 \). It follows that

\[
\ker(L_1) \oplus \cdots \oplus \ker(L_p) \subseteq \ker\left(\prod_{i=1}^{p} L_i\right)
\]

From Sylvester’s theorem and the observation about direct sums in Lemma \( \text{0.2} \),

\[
\sum_{i=1}^{p} \dim\left(\ker(L_i)\right) = \dim\left(\ker(L_1) \oplus \cdots \oplus \ker(L_p)\right)
\]

\[
\leq \dim\left(\ker\left(\prod_{i=1}^{p} L_i\right)\right) \leq \sum_{i=1}^{p} \dim\left(\ker(L_i)\right)
\]

which implies all these are equal. Now in general, if \( W \) is a subspace of \( V \), a finite dimensional vector space and the two have the same dimension, then \( W = V \), Lemma \( \text{6.3} \). It follows from * that

\[
\ker(L_1) \oplus \cdots \oplus \ker(L_p) = \ker\left(\prod_{i=1}^{p} L_i\right)
\]

So how does the above situation occur? First recall the following theorem and corollary about polynomials. It was Theorem \( \text{6.1.6} \) and Corollary \( \text{6.1.7} \) proved earlier.

**Theorem 6.1.6** Let \( f(\lambda) \) be a nonconstant polynomial with coefficients in \( \mathbb{F} \). Then there is some \( a \in \mathbb{F} \) such that \( f(\lambda) = a \prod_{i=1}^{p} \phi_i(\lambda) \) where \( \phi_i(\lambda) \) is an irreducible nonconstant monic polynomial and repeats are allowed. Furthermore, this factorization is unique in the sense that any two of these factorizations have the same nonconstant factors in the product, possibly in different order and the same constant \( a \).

**Corollary 6.1.7** Let \( q(\lambda) = \prod_{i=1}^{p} \phi_i(\lambda)^{k_i} \) where the \( k_i \) are positive integers and the \( \phi_i(\lambda) \) are irreducible monic polynomials. Suppose also that \( p(\lambda) \) is a monic polynomial which divides \( q(\lambda) \). Then

\[
p(\lambda) = \prod_{i=1}^{p} \phi_i(\lambda)^{r_i}
\]

where \( r_i \) is a nonnegative integer no larger than \( k_i \).
6.1. A THEOREM OF SYLVESTER, DIRECT SUMS

Now I will show how to use these basic theorems about polynomials to produce \( L_i \) such that the above major result follows. This is going to have a striking similarity to the notion of a minimum polynomial in the context of algebraic numbers.

**Definition 6.1.8** Let \( V \) be an \( n \) dimensional vector space, \( n \geq 1 \), and let \( L \in \mathcal{L}(V,V) \) which is a vector space of dimension \( n^2 \) by Theorem 6.1.4. Then \( p(\lambda) \) will be the non constant monic polynomial such that \( p(L) = 0 \) and out of all polynomials \( q(\lambda) \) such that \( q(L) = 0 \), the degree of \( p(\lambda) \) is the smallest. This is called the minimum polynomial. It is always understood that \( L \neq 0 \). It is not interesting to fuss with this case of the zero linear transformation.

In the following, we always define \( L^0 = I \).

**Theorem 6.1.9** The above definition is well defined. Also, if \( q(L) = 0 \), then \( p(\lambda) \) divides \( q(\lambda) \).

**Proof:** The dimension of \( \mathcal{L}(V,V) \) is \( n^2 \). Therefore, \( I, L, \ldots, L^{n^2} \) are linearly dependent and so there is some polynomial \( q(\lambda) \) such that \( q(L) = 0 \). Let \( m \) be the smallest degree of any polynomial with this property. Such a smallest number exists by well ordering of \( \mathbb{N} \). To obtain a monic polynomial \( p(\lambda) \) with degree \( m \), divide such a polynomial with degree \( m \), having the property that \( p(L) = 0 \) by the leading coefficient. Now suppose \( q(\lambda) \) is any polynomial such that \( q(L) = 0 \). Then by the Euclidean algorithm, there is \( r(\lambda) \) either zero or having degree less than the degree of \( p(\lambda) \) such that \( q(\lambda) = p(\lambda)k(\lambda) + r(\lambda) \) for some polynomial \( k(\lambda) \). But then

\[
0 = q(L) = k(L)p(L) + r(L)
\]

If \( r(\lambda) \neq 0 \), then this is a contradiction to \( p(\lambda) \) having the smallest degree. Therefore, \( p(\lambda) \) divides \( q(\lambda) \). Now suppose \( \hat{p}(\lambda) \) and \( p(\lambda) \) are two monic polynomials of degree \( m \). Then from what was just shown \( \hat{p}(\lambda) \) divides \( p(\lambda) \) and \( p(\lambda) \) divides \( \hat{p}(\lambda) \). Since they are both monic polynomials, they must be equal. Thus the minimum polynomial is unique and this shows the above definition is well defined.

Now here is the major result which comes from Sylvester’s theorem given above.

**Theorem 6.1.10** Let \( L \in \mathcal{L}(V,V) \) where \( V \) is an \( n \) dimensional vector space with field of scalars \( \mathbb{F} \). Letting \( p(\lambda) \) be the minimum polynomial for \( L \),

\[
p(\lambda) = \prod_{i=1}^{p} \phi_i(\lambda)^{k_i}
\]

where the \( k_i \) are positive integers and the \( \phi_i(\lambda) \) are irreducible monic polynomials. Also the linear maps \( \phi_i(L)^{k_i} \) commute and \( \phi_i(L)^{k_i} \) is one to one on \( \ker(\phi_j(L)^{k_j}) \) for all \( j \neq i \) as is \( \phi_i(L) \). Also

\[
V = \ker(\phi_1(L)^{k_1}) \oplus \cdots \oplus \ker(\phi_p(L)^{k_p})
\]

and each \( \ker(\phi_i(L)^{k_i}) \) is invariant with respect to \( L \). Letting \( L_j \) be the restriction of \( L \) to

\[
\ker(\phi_j(L)^{k_j}),
\]

it follows that the minimum polynomial of \( L_j \) equals \( \phi_j(\lambda)^{k_j} \). Also \( p \leq n \).

**Proof:** By Theorem 6.1.8 the minimum polynomial \( p(\lambda) \) is of the form

\[
a \prod_{i=1}^{p} \phi_i(\lambda)^{k_i}
\]
where \( \phi_i (\lambda) \) is monic and irreducible with \( \phi_i (\lambda) \neq \phi_j (\lambda) \) if \( i \neq j \). Since \( p (\lambda) \) is monic, it follows that \( a = 1 \). Since \( L \) commutes with itself, all of these \( \phi_i (L)^{k_i} \) commute. Also

\[
\phi_i (L) : \ker (\phi_j (L)^{k_j}) \to \ker (\phi_j (L)^{k_j})
\]

because all of these operators commute.

Now consider \( \phi_i (L) \). Is it one to one on \( \ker (\phi_j (L)^{k_j}) \)? Suppose not. Suppose that for some \( k \neq i, \phi_i (L) \) is not one to one on \( \ker (\phi_k (L)^{k_k}) \). We know that \( \phi_i (\lambda), \phi_k (\lambda)^{k_k} \) are relatively prime meaning the monic polynomial of greatest degree which divides them both is 1. Why is this? If some polynomial divided both, then it would need to be \( \phi_i (\lambda) \) or 1 because \( \phi_i (\lambda) \) is irreducible. But \( \phi_i (\lambda) \) cannot divide \( \phi_k (\lambda)^{k_k} \) unless it equals \( \phi_k (\lambda) \), this by Corollary \[\text{??}\] and they are assumed unequal. Hence there are polynomials \( l (\lambda), m (\lambda) \) such that

\[
1 = l (\lambda) \phi_i (\lambda) + m (\lambda) \phi_k (\lambda)^{k_k}
\]

By what we mean by equality of polynomials, that coefficients of equal powers of \( \lambda \) are equal, it follows that for \( I \) the identity transformation,

\[
I = l (\lambda) \phi_i (\lambda) + m (\lambda) \phi_k (\lambda)^{k_k}
\]

Say \( v \in \ker (\phi_k (L)^{k_k}) \) and \( v \neq 0 \) while \( \phi_i (L) v = 0 \). Then from the above equation,

\[
v = l (\lambda) \phi_i (\lambda) v + m (\lambda) \phi_k (\lambda)^{k_k} v = 0 + 0 = 0
\]

a contradiction. Thus \( \phi_i (L) \) and hence \( \phi_i (L)^{k_i} \) is one to one on \( \ker (\phi_j (L)^{k_j}) \). (Recall that, since these commute, \( \phi_i (L) \) maps \( \ker (\phi_j (L)^{k_j}) \) to \( \ker (\phi_i (L)^{k_i}) \).

Thus, from Lemma \[\text{??}\]

\[
V = \ker \left( \prod_{i=1}^{p} \phi_i (L)^{k_i} \right)
\]

\[
= \ker (\phi_1 (L)^{k_1}) \oplus \cdots \oplus \ker (\phi_p (L)^{k_p})
\]

Next consider the claim about the minimum polynomial of \( L_j \). Denote this minimum polynomial as \( p_j (\lambda) \). Then since \( \phi_j (L)^{k_j} = \phi_j (L_j)^{k_j} = 0 \) on \( \ker (\phi_j (L)^{k_j}) \), it must be the case that \( p_j (\lambda) \) must divide \( \phi_j (\lambda)^{k_j} \) and so by Corollary \[\text{??}\] this means \( p_j (\lambda) = \phi_j (\lambda)^{r_j} \) where \( r_j \leq k_j \). If \( r_j < k_j \), consider the polynomial

\[
\prod_{i=1, i \neq j}^{p} \phi_i (\lambda)^{k_i} \phi_j (\lambda)^{r_j} \equiv r (\lambda)
\]

Then since these operators \( \phi_i (L)^{k_i} \) commute with each other, \( r (L) = 0 \) because \( r (L) v = 0 \) for every \( v \in \ker (\phi_i (L)^{k_i}) \) and also \( r (L) v = 0 \) for \( v \in \ker (\phi_j (L)^{r_j}) \). However, this violates the definition of the minimum polynomial for \( L, p (\lambda) \) because here is a polynomial \( r (\lambda) \) such that \( r (L) = 0 \) but \( r (\lambda) \) has smaller degree than \( p (\lambda) \). Thus \( r_j = k_j \).

Consider the claim that \( p \leq n \) the dimension of \( V \). Let \( v_i \in \ker (\phi_i (L)^{k_i}) \), \( v_i \neq 0 \). Then it must be the case that \( \{v_1, \cdots, v_p\} \) is a linearly independent set because \( \ker (\phi_i (L)^{k_i}) \oplus \cdots \oplus \ker (\phi_p (L)^{k_p}) \) is a direct sum. Hence \( p \leq n \) because a linearly independent set is never longer than a spanning set one of which has \( n \) elements.
Letting $\beta_i$ be an ordered basis for $\ker(\phi_j(L)_{k_j})$ and letting $\beta \equiv (\beta_1, \beta_2, \cdots, \beta_p)$, it follows from Theorem 6.0.3, that if $M_j$ is the matrix for $L_j$, the restriction of $L$ to $\ker(\phi_j(L)_{k_j})$, then the matrix of $L$ with respect to the basis $\beta$ is a block diagonal matrix of the form

$$
\begin{pmatrix}
M_1 & 0 \\
\vdots & \ddots \\
0 & M_p
\end{pmatrix}
$$

The study of canonical forms has to do with choosing the bases $\beta_i$ in an auspicious manner. This topic will be discussed more later.

### 6.2 Finding The Minimum Polynomial

All of this depends on the minimum polynomial. It was shown above that this polynomial exists, but how can you find it? In fact, it is not all that hard to find. Recall that if $L \in \mathcal{L}(V, V)$ where the dimension of $V$ is $n$, then $I, L, L^2, \cdots, L^{n^2}$ is linearly independent. Thus some linear combination equals zero. The minimum polynomial was the polynomial $p(\lambda)$ of smallest degree which is monic and which has $p(L) = 0$. At this point, we only know that this degree is no more than $n^2$. However, it will be shown later in the proof of the Cayley Hamilton theorem that there exists a polynomial $q(\lambda)$ of degree $n$ such that $q(L) = 0$. Then from Theorem 6.1.9 it follows that $p(\lambda)$ divides $q(\lambda)$ and so the degree of $p(\lambda)$ will always be no more than $n$. Another observation to make is that it suffices to find the minimum polynomial for the matrix of the linear transformation taken with respect to any basis. Recall the relationship of this matrix and $L$.

$$
\begin{align*}
L & \colon V \to V \\
q_i & \colon q_i \to q_i \\
F^n & \to F^n \\
A & \colon A
\end{align*}
$$

where $q$ is a one to one and onto linear map from $F^n$ to $V$. Thus if $p(L)$ is a polynomial in $L$,

$$
p(L) = p(q^{-1}Aq)
$$

A typical term on the right is of the form

$$
c_k \left( \underbrace{q^{-1}Aq \cdots q^{-1}Aq}_k \right) = q^{-1}(c_kA^k)q
$$

Thus, applying this to each term and factoring out $q^{-1}$ and $q$,

$$
p(L) = q^{-1}p(A)q
$$

Recall the convention that $A^0 = I$ the identity matrix and $L^0 = I$, the identity linear transformation. Thus $p(L) = 0$ if and only if $p(A) = 0$ and so the minimum polynomial for $A$ is exactly the same as the minimum polynomial for $L$. However, in case of $A$, the multiplication is just matrix multiplication so we can compute with it easily.

This shows that it suffices to learn how to find the minimum polynomial for an $n \times n$ matrix. I will show how to do this with some examples. The process can be made much more systematic, but I will try to keep it pretty short because it is often the case that it is easy to find it without going through a long computation.
Example 6.2.1 Find the minimum polynomial of

\[
\begin{pmatrix}
-1 & 0 & 6 \\
1 & 1 & -3 \\
-1 & 0 & 4
\end{pmatrix}
\]

Go right to the definition and use the fact that you only need to have three powers of this matrix in order to get things to work. Thus the minimum polynomial involves finding \( a, b, c, d \) scalars such that

\[
a \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix} + b \begin{pmatrix}
-1 & 0 & 6 \\
1 & 1 & -3 \\
-1 & 0 & 4
\end{pmatrix} + c \begin{pmatrix}
-1 & 0 & 6 \\
1 & 1 & -3 \\
-1 & 0 & 4
\end{pmatrix}^2 + d \begin{pmatrix}
-1 & 0 & 6 \\
1 & 1 & -3 \\
-1 & 0 & 4
\end{pmatrix}^3 = 0
\]

You could include all nine powers if you want, but there is no point in doing so from what will be presented later.

There is such a solution from the above theory and it is only a matter of finding it. Thus you need to find scalars such that

\[
a \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix} + b \begin{pmatrix}
-1 & 0 & 6 \\
1 & 1 & -3 \\
-1 & 0 & 4
\end{pmatrix} + c \begin{pmatrix}
-5 & 0 & 18 \\
3 & 1 & -9 \\
-3 & 0 & 10
\end{pmatrix} + d \begin{pmatrix}
-13 & 0 & 42 \\
7 & 1 & -21 \\
-7 & 0 & 22
\end{pmatrix} = 0
\]

Lets try the diagonal entries first and then lets pick the bottom left corner.

\[
a - b - 5c - 13d = 0 \\
a + b + c + d = 0 \\
a + 4b + 10c + 22d = 0 \\
-b + 3c + 7d = 0
\]

Thus we row reduce the matrix

\[
\begin{pmatrix}
1 & -1 & -5 & -13 & 0 \\
1 & 1 & 1 & 1 & 0 \\
1 & 4 & 10 & 22 & 0 \\
0 & -1 & -3 & -7 & 0
\end{pmatrix}
\]

which yields after some computations

\[
\begin{pmatrix}
1 & 0 & -2 & -6 & 0 \\
0 & 1 & 3 & 7 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

We can take \( d = 0 \) and \( c = 1 \) and find that \( a = 2, b = -3 \). Thus a candidate for minimum polynomial is

\[
\lambda^2 - 3\lambda + 2.
\]

Could you have a smaller degree polynomial? No you could not because if you took both \( c \) and \( d \) equal to 0, then you would be forced to have \( a, b \) both be zero as well. Hence this must be the minimum polynomial \textbf{provided the matrix satisfies this equation}. However, you can just plug it in to the equation and see that it works. If it didn’t work, you would simply include another equation in the above computation for \( a, b, c, \) and \( d \).

\[
\begin{pmatrix}
-1 & 0 & 6 \\
1 & 1 & -3 \\
-1 & 0 & 4
\end{pmatrix}^2 - 3 \begin{pmatrix}
-1 & 0 & 6 \\
1 & 1 & -3 \\
-1 & 0 & 4
\end{pmatrix} + 2 \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix} = \begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
\]
6.3. EIGENVALUES AND EIGENVECTORS OF LINEAR TRANSFORMATIONS

It is a little tedious, but completely routine to find this minimum polynomial. To be more systematic, you would take the powers of the matrix and string each of them out into a long $n^2 \times 1$ vector and make these the columns of a matrix which would then be row reduced. However, as shown above, you can get away with less as in the above example, but you need to be sure to check that the matrix satisfies the equation you come up with.

Now here is an example where $\mathbb{F} = \mathbb{Z}_5$ and the arithmetic is in $\mathbb{F}$.

**Example 6.2.2** The matrix is

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 3 & 1 \\ 4 & 1 & 1 \end{pmatrix}$$

Find the minimum polynomial.

Powers of the matrix are

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 0 & 3 & 1 \\ 4 & 1 & 1 \end{pmatrix}, \begin{pmatrix} 3 & 1 & 3 \\ 4 & 0 & 4 \\ 3 & 2 & 4 \end{pmatrix}, \begin{pmatrix} 0 & 2 & 3 \\ 4 & 1 & 0 \end{pmatrix}$$

Then we look for the polynomial of smallest degree such that $a + b\lambda + c\lambda^2 + d\lambda^3 = 0$. An appropriate augmented matrix is

$$\begin{pmatrix} 1 & 1 & 3 & 0 & 0 \\ 1 & 3 & 0 & 2 & 0 \\ 1 & 1 & 4 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \end{pmatrix}$$

Then row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

so it would seem a possible minimum polynomial is obtained by $a = 1, b = -1 = 4, c = 0, d = 1$. Thus it has degree 3. There cannot be any polynomial of smaller degree because of the first three columns so it would seem that this should be the minimum polynomial,

$$1 + 4\lambda + \lambda^3$$

Does it send the matrix to 0? This just involves checking whether it does and in fact, this is the case using the arithmetic in the residue class.

6.3 Eigenvalues And Eigenvectors Of Linear Transformations

We begin with the following fundamental definition.

**Definition 6.3.1** Let $L \in \mathcal{L}(V, V)$ where $V$ is a vector space of dimension $n$ with field of scalars $\mathbb{F}$. An eigen-pair consists of a scalar $\lambda \in \mathbb{F}$ called an eigenvalue and a \textbf{non-zero} $v \in V$ such that

$$(\lambda I - L)v = 0$$
Do eigen-pairs exist? Recall that from Theorem 6.3.1 the minimum polynomial can be factored in a unique way as
\[ p(\lambda) = \prod_{i=1}^{p} (\lambda - \mu_i)^{k_i} \]
where each \( \phi_i(\lambda) \) is irreducible and monic. Then the following theorem is obtained.

**Theorem 6.3.2** Suppose the minimum polynomial of \( L \in \mathcal{L}(V, V) \), \( p(\lambda) \) factors completely into linear factors (splits) so that
\[ p(\lambda) = \prod_{i=1}^{p} (\lambda - \mu_i)^{k_i} \]

Then the \( \mu_i \) are eigenvalues and corresponding to each of these eigenvalues, there is an eigenvector \( w_i \neq 0 \) such that \( Lw_i = \mu_i w_i \). Also, there are no other eigenvectors than these \( \mu_i \). Also \( V = \ker(L - \mu_1 I)^{k_1} \oplus \cdots \oplus \ker(L - \mu_p I)^{k_p} \)

and if \( L_i \) is the restriction of \( L \) to \( \ker(L - \mu_i I)^{k_i} \), then \( L_i \) has exactly one eigenvalue and it is \( \mu_i \).

**Proof:** In this case, the irreducible polynomials are of the form \( \phi_i(\lambda) = \lambda - \mu_i \) so the claim about the form of \( p(\lambda) \) is valid. Now by the definition of \( p(\lambda) \) as the unique polynomial having smallest degree for which \( p(\lambda) = 0 \), there must exist some vector \( v_j \neq 0 \) such that
\[ \prod_{i=1, i \neq j}^{p} (L - \mu_i I)^{k_i} (L - \mu_j I)^{k_j - 1} v_j \neq 0 \]

This is because \( \prod_{i=1, i \neq j}^{p} (L - \mu_i I)^{k_i} (L - \mu_j I)^{k_j} \) has smaller degree than \( p(\lambda) \) and so it is not zero. But
\[ (L - \mu_j I) \prod_{i=1, i \neq j}^{p} (L - \mu_i I)^{k_i} (L - \mu_j I)^{k_j - 1} v_j = \prod_{i=1}^{p} (\lambda - \mu_i)^{k_i} v_j = 0 \]

and so
\[ w_j \equiv \prod_{i=1, i \neq j}^{p} (L - \mu_i I)^{k_i} (L - \mu_j I)^{k_j - 1} v_j \]

must be an eigenvector. Thus each of these \( (\mu_j, w_j) \) is a eigen-pair.

Now suppose \( \mu \) is an eigenvector. Why is it one of the \( \mu_i \)? Say \( (L - \mu I) w = 0 \) for \( w \neq 0 \) and suppose \( \mu \) is not one of the \( \mu_i \). Then using the binomial theorem as required,
\[ 0 = \prod_{i=1}^{p} (L - \mu_i I)^{k_i} w = \prod_{i=1}^{p} (L - \mu I + (\mu - \mu_i) I)^{k_i} w \]
\[ = \prod_{i=1}^{p} \sum_{s=0}^{k_i} \binom{k_i}{s} (\mu - \mu_i)^{s} (L - \mu I)^{k_i - s} w = \prod_{i=1}^{p} (\mu - \mu_i)^{k_i} w \]

which is impossible if none of the \( \mu_i \) equal \( \mu \).

Consider the last assertion about \( L_i \). From Theorem 6.3.1 the minimum polynomial of \( L_i \) is \( (\lambda - \mu_i I)^{k_i} \). Therefore, there exists a vector \( x \in \ker(L_i - \mu_i I)^{k_i} = \ker(L - \mu_i I)^{k_i} \) such that \( (L_i - \mu_i I)^{k_i - 1} x \neq 0 \). Otherwise, this would not really be the minimum polynomial. Then
\[ (L_i - \mu_i I) (L_i - \mu_i I)^{k_i - 1} x = 0 \]

and so \( \mu_i \) is indeed an eigenvalue of \( L_i \). Could \( L_i \) have another eigenvalue \( \mu \)? No, it could not. This is by a repeat of the above argument in the simpler case where \( p = 1 \).
Example 6.3.3  The minimum polynomial for the matrix

\[ A = \begin{pmatrix} 4 & 0 & -6 \\ -1 & 2 & 3 \\ 1 & 0 & -1 \end{pmatrix} \]

is \(\lambda^2 - 3\lambda + 2\). This factors as \((\lambda - 2)(\lambda - 1)\) and so the eigenvalues are 1, 2. Find the eigen-pairs. Then determine the matrix with respect to a basis of these eigenvectors if possible.

First consider the eigenvalue 2. There exists a nonzero vector \(v\) such that \((A - 2I)v = 0\). This follows from the above theory. However, it is best to just find it directly rather than try to get it by using the proof of the above theorem. The augmented matrix to consider is then

\[
\begin{pmatrix} 4-2 & 0 & -6 & 0 \\ -1 & 2-2 & 3 & 0 \\ 1 & 0 & -1-2 & 0 \end{pmatrix}
\]

Row reducing this yields

\[
\begin{pmatrix} 1 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}
\]

Thus the solution is any vector of the form

\[
\begin{pmatrix} 3z \\ y \\ z \end{pmatrix} = z \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \; z, y \text{ not both 0}
\]

Now consider the eigenvalue 1. This time you row reduce

\[
\begin{pmatrix} 4-1 & 0 & -6 & 0 \\ -1 & 2-1 & 3 & 0 \\ 1 & 0 & -1-1 & 0 \end{pmatrix}
\]

which yields for the row reduced echelon form

\[
\begin{pmatrix} 1 & 0 & -2 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}
\]

Thus an eigenvector is of the form

\[
\begin{pmatrix} 2z \\ -z \\ z \end{pmatrix}, \; z \neq 0
\]

Consider a basis for \(\mathbb{R}^n\) of the form

\[
\left\{ \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \right\}
\]

You might want to consider Problem 9 on Page 113 at this point. This problem shows that the matrix with respect to this basis is diagonal.
When the matrix of a linear transformation can be chosen to be a diagonal matrix, the transformation is said to be nondefective. Also, note that the term applies to the matrix of a linear transformation and so I will specialize to the consideration of matrices in what follows. As shown above, this is equivalent to saying that any matrix of the linear transformation is similar to one which is diagonal. That is, the matrix of a linear transformation, or more generally just a square matrix \( A \) has the property that there exists \( S \) such that
\[
S^{-1}AS = D
\]
where \( D \) is a diagonal matrix. That is, the matrix of a linear transformation, or more generally just a square matrix
\[
A
\]
has the property that there exists \( S \) such that
\[
S^{-1}AS = D
\]
where \( D \) is a diagonal matrix. This means \( D \) has a zero as every entry except for the main diagonal. More precisely, \( D_{ij} = 0 \) unless \( i = j \). Such matrices look like the following.
\[
\begin{pmatrix}
* & 0 \\
. & . \\
0 & *
\end{pmatrix}
\]
where \( * \) might not be zero.

The most important theorem about diagonalizability is the following major result. First here is a simple observation.

**Observation 6.3.5** Let \( S = \begin{pmatrix} s_1 & \cdots & s_n \end{pmatrix} \) where \( S \) is \( n \times n \). Then here is the result of multiplying on the right by a diagonal matrix.
\[
\begin{pmatrix} s_1 & \cdots & s_n \end{pmatrix} \begin{pmatrix} \lambda_1 \\
. \\
. \\
\lambda_n
\end{pmatrix} = \begin{pmatrix} \lambda_1 s_1 & \cdots & \lambda_n s_n \end{pmatrix}
\]
This follows from the way we multiply matrices. The \( i \)th entry of the \( j \)th column of the product on the left is of the form \( s_i \lambda_j \). Thus the \( j \)th column of the matrix on the left is just \( \lambda_j s_j \).

**Theorem 6.3.6** An \( n \times n \) matrix is diagonalizable if and only if \( \mathbb{F}^n \) has a basis of eigenvectors of \( A \). Furthermore, you can take the matrix \( S \) described above, to be given as
\[
S = \begin{pmatrix} s_1 & s_2 & \cdots & s_n \end{pmatrix}
\]
where here the \( s_k \) are the eigenvectors in the basis for \( \mathbb{F}^n \). If \( A \) is diagonalizable, the eigenvalues of \( A \) are the diagonal entries of the diagonal matrix.

**Proof:** To say that \( A \) is diagonalizable, is to say that
\[
S^{-1}AS = \begin{pmatrix} \lambda_1 \\
. \\
. \\
\lambda_n
\end{pmatrix}
\]
the \( \lambda_i \) being elements of \( \mathbb{F} \). This is to say that for \( S = \begin{pmatrix} s_1 & \cdots & s_n \end{pmatrix} \), \( s_k \) being the \( k \)th column,
\[
A \begin{pmatrix} s_1 & \cdots & s_n \end{pmatrix} = \begin{pmatrix} s_1 & \cdots & s_n \end{pmatrix} \begin{pmatrix} \lambda_1 \\
. \\
. \\
\lambda_n
\end{pmatrix}
\]

---

\(^1\)This word has 9 syllables! Such words belong in Iceland. Eyjafjallajökull actually only has seven syllables.
which is equivalent, from the way we multiply matrices, that

\[
\begin{pmatrix}
    A s_1 & \cdots & A s_n
\end{pmatrix} =
\begin{pmatrix}
    \lambda_1 s_1 & \cdots & \lambda_n s_n
\end{pmatrix}
\]

which is equivalent to saying that the columns of \( S \) are eigenvectors and the diagonal matrix has the eigenvectors down the main diagonal. Since \( S^{-1} \) is invertible, these eigenvectors are a basis. Similarly, if there is a basis of eigenvectors, one can take them as the columns of \( S \) and reverse the above steps, finally concluding that \( A \) is diagonalizable.

**Corollary 6.3.7** Let \( A \) be an \( n \times n \) matrix with minimum polynomial

\[
p(\lambda) = \prod_{i=1}^{p} (\lambda - \mu_i)^{k_i}, \text{ the } \mu_i \text{ being distinct}.
\]

Then \( A \) is diagonalizable if and only if each \( k_i = 1 \).

**Proof:** Suppose first that it is diagonalizable and that a basis of eigenvectors is \( \{v_1, \ldots, v_n\} \) with \( Av_i = \mu_i v_i \). Since \( n \geq p \), there may be some repeats here, a \( \mu_i \) going with more than one \( v_i \). Say \( k_i > 1 \). Now consider \( \hat{p}(\lambda) = \prod_{j=1, j \neq i}^{p} (\lambda - \mu_j)^{k_j} (\lambda - \mu_i) \). Thus this is a monic polynomial which has smaller degree than \( p(\lambda) \). If you have \( v \in \mathbb{F}^n \), since this is a basis, there are scalars \( c_i \) such that \( v = \sum_j c_j v_j \). Then \( \hat{p}(A)v = 0 \). Since \( v \) is arbitrary, this shows that \( \hat{p}(A) = 0 \) contrary to the definition of the minimum polynomial being \( p(\lambda) \). Thus each \( k_i \) must be 1.

Conversely, if each \( k_i = 1 \), then

\[
\mathbb{F}^n = \ker (A - \mu_1 I) \oplus \cdots \oplus \ker (A - \mu_p I)
\]

and you simply let \( \beta_i \) be a basis for \( \ker (A - \mu_i I) \) which consists entirely of eigenvectors by definition of what you mean by \( \ker (A - \mu_i I) \). Then a basis of eigenvectors consists of \( \{\beta_1, \beta_2, \ldots, \beta_p\} \) and so the matrix \( A \) is diagonalizable.

**Example 6.3.8** The minimum polynomial for the matrix

\[
A = \begin{pmatrix}
    10 & 12 & -6 \\
    -4 & -4 & 3 \\
    3 & 4 & -1
\end{pmatrix}
\]

is \( \lambda^3 - 5\lambda^2 + 8\lambda - 4 \). This factors as \((\lambda - 2)^2(\lambda - 1)\) and so the eigenvalues are 1, 2. Find the eigen-pairs. Then determine the matrix with respect to a basis of these eigenvectors if possible. If it is not possible to find a basis of eigenvectors, find a block diagonal matrix similar to the matrix.

First find the eigenvectors for 2. You need to row reduce

\[
\begin{pmatrix}
    10 - 2 & 12 & -6 & 0 \\
    -4 & -4 - 2 & 3 & 0 \\
    3 & 4 & -1 - 2 & 0
\end{pmatrix}
\]

This yields

\[
\begin{pmatrix}
    1 & 0 & -3 & 0 \\
    0 & 1 & \frac{3}{2} & 0 \\
    0 & 0 & 0 & 0
\end{pmatrix}
\]

Thus the eigenvectors which go with 2 are

\[
\begin{pmatrix}
    6z \\
    -3z \\
    2z
\end{pmatrix}, \quad z \in \mathbb{R}, \quad z \neq 0
\]
The eigenvectors which go with 1 are
\[ z \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}, \quad z \in \mathbb{R}, \quad z \neq 0 \]

By Theorem 6.3.2, there are no other eigenvectors than those which correspond to eigenvalues 1, 2. Thus there is no basis of eigenvectors because the span of the eigenvectors has dimension two.

However, we can consider
\[ \mathbb{R}^3 = \ker \left( (A - 2I)^2 \right) \oplus \ker (A - I) \]

The second of these is just span \((\begin{pmatrix} 2 & -1 & 1 \end{pmatrix})^T\). What is the first? We find it by row reducing the following matrix which is the square of \(A - 2I\) augmented with a column of zeros.

\[
\begin{pmatrix}
-2 & 0 & 6 & 0 \\
1 & 0 & -3 & 0 \\
-1 & 0 & 3 & 0
\end{pmatrix}
\]

Row reducing this yields
\[
\begin{pmatrix}
1 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]

which says that solutions are of the form
\[
\begin{pmatrix}
3z \\
y \\
z
\end{pmatrix}, \quad y, z \in \mathbb{R} \text{ not both 0}
\]

This is the nonzero vectors of
\[ \text{span} \left( \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right) \]

What is the matrix of the restriction of \(A\) to this subspace?

\[
A \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 10 & 12 & -6 \\ -4 & -4 & 3 \\ 3 & 4 & -1 \end{pmatrix} \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 24 \\ -9 \\ 8 \end{pmatrix}
\]

\[
A \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 10 & 12 & -6 \\ -4 & -4 & 3 \\ 3 & 4 & -1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 12 \\ -4 \\ 4 \end{pmatrix}
\]

Then
\[
\begin{pmatrix}
24 & 12 \\
-9 & -4 \\
8 & 4
\end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} M \quad (6.2)
\]

and so some computations yield
\[
M = \begin{pmatrix} 8 & 4 \\ -9 & -4 \end{pmatrix}
\]
Indeed this works

\[
\begin{pmatrix}
3 & 0 \\
0 & 1 \\
1 & 0 \\
\end{pmatrix}
\begin{pmatrix}
8 & 4 \\
-9 & -4 \\
\end{pmatrix}
= 
\begin{pmatrix}
24 & 12 \\
-9 & -4 \\
8 & 4 \\
\end{pmatrix}
\]

Then the matrix associated with the other eigenvector is just 1. Hence the matrix with respect to the above ordered basis is

\[
\begin{pmatrix}
8 & 4 & 0 \\
-9 & -4 & 0 \\
0 & 0 & 1 \\
\end{pmatrix}
\]

So what are some convenient computations which will allow you to find \(M\) easily? Take the transpose of both sides of 6.2. Then you would have

\[
\begin{pmatrix}
24 & -9 & 8 \\
12 & -4 & 4 \\
\end{pmatrix}
= 
M^T
\begin{pmatrix}
3 & 0 & 1 \\
0 & 1 & 0 \\
\end{pmatrix}
\]

Thus

\[
M^T
\begin{pmatrix}
0 \\
1 \\
\end{pmatrix}
= 
\begin{pmatrix}
-9 \\
-4 \\
\end{pmatrix}
, 
M^T
\begin{pmatrix}
1 \\
0 \\
\end{pmatrix}
= 
\begin{pmatrix}
8 \\
4 \\
\end{pmatrix}
\]

and so \(M^T = \begin{pmatrix} 8 & -9 \\ 4 & -4 \end{pmatrix}\) so \(M = \begin{pmatrix} 8 & 4 \\ -9 & -4 \end{pmatrix}\).

In these examples given above, it was possible to factor the minimum polynomial and explicitly determine eigenvalues and eigenvectors and obtain information about whether the matrix was diagonalizable by explicit computations. Well, what if you can’t factor the minimum polynomial? What then? This is the typical situation, not what was presented in the above examples. Just write down a 3 \(\times\) 3 matrix and see if you can find the eigenvalues explicitly using algebra. Is there a way to determine whether a given matrix is diagonalizable in the case that the minimum polynomial factors although you might have trouble finding the factors? Amazingly, the answer is yes. One can answer this question completely using only methods from algebra.

### 6.4 A Formal Derivative, Diagonalizability

For \(p(\lambda) = a_n\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0\) where \(n\) is a positive integer, define

\[p'(\lambda) \equiv na_n\lambda^{n-1} + (n-1)a_{n-1}\lambda^{n-2} + \cdots + a_1\]

In other words, you use the usual rules of differentiation in calculus to write down this formal derivative. It has absolutely no physical significance in this context because the coefficients are just elements of some field, possibly \(\mathbb{Z}_p\). It is a purely algebraic manipulation. A term like \(ka\) where \(k \in \mathbb{N}\) and \(a \in \mathbb{F}\) means to add \(a\) to itself \(k\) times. There are no limits or anything else. However, this has certain properties. In particular, the “derivative” of a sum equals the sum of the derivatives. Also

\[
(b\lambda^m \cdot (a_n\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0))' = 
(a_n b\lambda^{n+m} + ba_{n-1}\lambda^{m+(n-1)} + \cdots + ba_1\lambda^{1+m} + a_0 b\lambda^m)' = 
a_n b (n+m)\lambda^{n+m-1} + ba_{n-1} (m+n-1)\lambda^{n+m-2} + \cdots + ba_1 (m+1)\lambda^m + a_0 bm\lambda^{m-1}
\]

Will the product rule give the same thing? Is it true that the above equals

\[
(b\lambda^m)' (a_n\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0) + 
+ b\lambda^m (a_n\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0)'?
\]
A short computation shows that this is indeed the case. Then by induction one can conclude that

$$\left(\prod_{i=1}^{p} p_i(\lambda)\right) = \sum_{j=1}^{p} p'_j(\lambda) \prod_{i \neq j} p_i(\lambda)$$

In particular, if

$$p(\lambda) = \prod_{i=1}^{p} (\lambda - \mu_i)^{k_i}$$

then

$$p'(\lambda) = \sum_{j=1}^{p} k_j (\lambda - \mu_j)^{k_j-1} \prod_{i \neq j} (\lambda - \mu_i)^{k_i}$$

I want to emphasize that this is an arbitrary field of scalars, but if one is only interested in the real or complex numbers, then all of this follows from standard calculus theorems.

**Proposition 6.4.1** Suppose the minimum polynomial \( p(\lambda) \) of an \( n \times n \) matrix \( A \) completely factors into linear factors. Then \( A \) is diagonalizable if and only if \( p(\lambda), p'(\lambda) \) are relatively prime.

**Proof:** Suppose \( p(\lambda), p'(\lambda) \) are relatively prime. Say

$$p(\lambda) = \prod_{i=1}^{n} (\lambda - \mu_i)^{k_i}, \ \mu_i \text{ are distinct}$$

From the above discussion,

$$p'(\lambda) = \sum_{j=1}^{p} k_j (\lambda - \mu_j)^{k_j-1} \prod_{i \neq j} (\lambda - \mu_i)^{k_i}$$

and \( p'(\lambda), p(\lambda) \) are relatively prime if and only if each \( k_i = 1 \). Then by Corollary 6.3.7 this is true if and only if \( A \) is diagonalizable. \( \blacksquare \)

**Example 6.4.2** Find whether the matrix

$$A = \begin{pmatrix} 1 & -1 & 2 \\ 0 & 1 & 2 \\ 1 & -1 & 1 \end{pmatrix}$$

is diagonalizable. Assume the field of scalars is \( \mathbb{C} \) because in this field, the minimum polynomial will factor thanks to the fundamental theorem of algebra.

Successive powers of the matrix are

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & -1 & 2 \\ 0 & 1 & 2 \\ 1 & -1 & 1 \end{pmatrix}, \begin{pmatrix} 3 & -4 & 2 \\ 2 & -1 & 4 \\ 2 & -3 & 1 \end{pmatrix}, \begin{pmatrix} 5 & -9 & 0 \\ 6 & -7 & 6 \\ 3 & -6 & -1 \end{pmatrix}$$

Then we need to have for a linear combination involving \( a, b, c, d \) as scalars

$$a + b + 3c + 5d = 0$$
$$2c + 6d = 0$$
$$b + 2c + 3d = 0$$
$$-b - 3c - 6d = 0$$
Then letting \( d = 1 \), this gives only one solution,

\[
a = 1, b = 3, c = -3
\]

and so the candidate for the minimum polynomial is

\[
\lambda^3 - 3\lambda^2 + 3\lambda + 1
\]

In fact, this does work as is seen by substituting \( A \) for \( \lambda \). So is this polynomial and its derivative relatively prime?

\[
\lambda^3 - 3\lambda^2 + 3\lambda + 1 = \frac{1}{3}(\lambda - 1)(3\lambda^2 - 6\lambda + 3) + 2
\]

and clearly \((3\lambda^2 - 6\lambda + 3)\) and 2 are relatively prime. Hence this matrix is diagonalizable. Of course, finding its diagonalization is another matter. For this algorithm for determining whether two polynomials are relatively prime, see Problem 34 on Page 31.

Of course this was an easy example thanks to Problem 12 on Page 142, because there are three distinct eigenvalues, one real and two complex which must be complex conjugates. This problem says that eigenvectors corresponding to distinct eigenvalues are an independent set.

Consider the following example in which the eigenvalues are not distinct, consisting of \( a, a \).

**Example 6.4.3** Find whether the matrix

\[
A = \begin{pmatrix}
a + 1 & 1 \\
-1 & a - 1
\end{pmatrix}
\]

is diagonalizable.

Listing the powers of the matrix,

\[
\begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}, \begin{pmatrix}a + 1 & 1 \\ -1 & a - 1\end{pmatrix}, \begin{pmatrix}a^2 + 2a & 2a \\ -2a & a^2 - 2a\end{pmatrix}
\]

Then we need to have for a linear combination involving scalars \( x, y, z \)

\[
x + (a + 1)y + (a^2 + 2a)z = 0
\]

\[
y + 2az = 0
\]

\[
x + (a - 1)y + (a^2 - 2a)z = 0
\]

Then some routine row operations yield \( x = a^2z \), \( y = -2az \) and \( z \) is arbitrary. For the minimum polynomial, we take \( z = 1 \) because this is a monic polynomial. Thus the minimum polynomial is

\[
a^2 - 2a\lambda + \lambda^2 = (\lambda - a)^2
\]

and clearly this and its derivative are not relatively prime. Thus this matrix is not diagonalizable for any choice of \( a \).

### 6.5 Exercises

1. For the linear transformation on \( \mathbb{R}^2 \) determined by multiplication by the following matrices, find the minimum polynomial.

   \[
   \begin{pmatrix}
   3 & 1 \\
   -4 & -1
   \end{pmatrix}
   \]

   \[
   \begin{pmatrix}
   0 & -2 \\
   1 & 3
   \end{pmatrix}
   \]
2. Here is a matrix:
\[
\begin{pmatrix}
-1 & -4 & -2 \\
2 & 4 & 2 \\
-1 & -1 & 0
\end{pmatrix}
\]
Its minimum polynomial is \(\lambda^3 - 3\lambda^2 + 4\lambda - 2 = (\lambda - 1)(-2\lambda + \lambda^2 + 2)\). Obtain a block diagonal matrix similar to this one.

3. Suppose \(A \in \mathcal{L}(V, V)\) where \(V\) is a finite dimensional vector space and suppose \(p(\lambda)\) is the minimum polynomial. Say \(p(\lambda) = \lambda^m + a_{m-1}\lambda^{m-1} + \cdots + a_1\lambda + a_0\). If \(A\) is one to one, show that it is onto and also that \(A^{-1} \in \mathcal{L}(V, V)\). In this case, explain why \(a_0 \neq 0\). In this case, give a formula for \(A^{-1}\) as a polynomial in \(A\).

4. Let \(A = \begin{pmatrix} 0 & -2 \\ 1 & 3 \end{pmatrix}\). Its minimum polynomial is \(\lambda^2 - 3\lambda + 2\). Find \(A^{10}\) exactly. **Hint:** You can do long division and get \(\lambda^{10} = l(\lambda)(\lambda^2 - 3\lambda + 2) + 1023\lambda - 1022\).

5. Suppose \(A \in \mathcal{L}(V, V)\) and it has minimum polynomial \(p(\lambda)\) which has degree \(m\). It is desired to compute \(A^n\) for \(n\) large. Show that it is possible to obtain \(A^n\) in terms of a polynomial in \(A\) of degree less than \(m\).

6. Determine whether the following matrices are diagonalizable. Assume the field of scalars is \(\mathbb{C}\).

(a) \(
\begin{pmatrix}
1 & 1 & 1 \\
-1 & 2 & 1 \\
0 & 1 & 1
\end{pmatrix}
\)
(b) \(
\begin{pmatrix}
\sqrt{2} + 1 & 1 \\
-1 & \sqrt{2} - 1
\end{pmatrix}
\)
(c) \(
\begin{pmatrix}
a + 1 & 1 \\
-1 & a - 1
\end{pmatrix}
\) where \(a\) is a real number.
(d) \(
\begin{pmatrix}
1 & 1 & -1 \\
2 & 1 & -1 \\
0 & 1 & 2
\end{pmatrix}
\)
6.5. Exercises

(e) \[
\begin{pmatrix}
2 & 1 & 0 \\
-1 & 0 & 0 \\
2 & 2 & 1
\end{pmatrix}
\]

7. The situation for diagonalizability was presented for the situation in which the minimum polynomial factors completely as a product of linear factors since this is certainly the case of most interest, including \( \mathbb{C} \). What if the minimum polynomial does not split? Is there a theorem available that will allow one to conclude that the matrix is diagonalizable in a splitting field, possibly larger than the given field? It is a reasonable question because the assumption that \( p(\lambda), p'(\lambda) \) are relatively prime may be determined without factoring the polynomials and involves only computations involving the given field \( \mathbb{F} \). If you enlarge the field, what happens to the minimum polynomial? Does it stay the same or does it change? Remember, the matrix has entries all in the smaller field \( \mathbb{F} \) while a splitting field is \( \mathbb{G} \) larger than \( \mathbb{F} \), but you can determine the minimum polynomial using row operations on vectors in \( \mathbb{F}^n \).

8. Suppose \( V \) is a finite dimensional vector space and suppose \( N \in \mathcal{L}(V, V) \) satisfies \( N^m = 0 \) for some \( m \geq 1 \). Show that the only eigenvalue is 0.

9. Suppose \( V \) is an \( n \) dimensional vector space and suppose \( \beta \) is a basis for \( V \). Consider the map \( \mu I : V \to V \) given by \( \mu I v = \mu v \). What is the matrix of this map with respect to the basis \( \beta \)?

**Hint:** You should find that it is \( \mu \) times the identity matrix whose \( ij^{th} \) entry is \( \delta_{ij} \) which is 1 if \( i = j \) and 0 if \( i \neq j \). Thus the \( ij^{th} \) entry of this matrix will be \( \mu \delta_{ij} \).

10. In the case that the minimum polynomial factors, which was discussed above, we had

\[
V = \ker (L - \mu_1 I)^{k_1} \oplus \cdots \oplus \ker (L - \mu_p I)^{k_p}
\]

If \( V_i = \ker (L - \mu_i I)^{k_i} \), then by definition, \( (L - \mu_i I)^{k_i} = 0 \) where here \( L_i \) is the restriction of \( L \) to \( V_i \). If \( N = L_i - \mu_i I \), then \( N : V_i \to V_i \) and \( N^{k_i} = 0 \). This is the definition of a nilpotent transformation, one which has a high enough power equal to 0. Suppose then that \( N : V \to V \) where \( V \) is an \( m \) dimensional vector space. We will show that there is a basis for \( V \) such that with respect to this basis, the matrix of \( N \) is block diagonal and of the form

\[
\begin{pmatrix}
N_1 & 0 \\
\vdots & \ddots \\
0 & N_s
\end{pmatrix}
\]

where \( N_i \) is an \( r_i \times r_i \) matrix of the form

\[
\begin{pmatrix}
0 & 1 & 0 \\
0 & \ddots & 0 \\
0 & \cdots & 1 \\
0 & 0 & 0
\end{pmatrix}
\]

That is, there are ones down the superdiagonal and zeros everywhere else. Now consider the case where \( N_i = L_i - \mu_i I \) on one of the \( V_i \) as just described. Use the preceding problem and the special basis \( \beta_i \) just described for \( N_i \) to show that the matrix of \( L_i \) with respect to this basis is of the form

\[
J(\mu_i) \equiv \begin{pmatrix}
J_1(\mu_i) & 0 \\
\vdots & \ddots \\
0 & J_s(\mu_i)
\end{pmatrix}
\]
Now in this problem, the method for finding the special basis for a nilpotent transformation
Show that the vectors in any chain are linearly independent and for

Consider the set of all chains based on eigenvectors. Since all have total length no larger

This special matrix is called the Jordan canonical form. This problem shows that it reduces to
the study of the matrix of a nilpotent matrix. You see that it is a block diagonal matrix such
that each block is a block diagonal matrix which is also an upper triangular matrix having the
eigenvalues down the main diagonal and strings of ones on the super diagonal.

11. Now in this problem, the method for finding the special basis for a nilpotent transformation
is given. Let \( V \) be a vector space and let \( N \in \mathcal{L}(V, V) \) be nilpotent. First note the only
eigenvalue of \( N \) is 0. Why? (See Problem 3.) Let \( v_1 \) be an eigenvector. Then \( \{v_1, v_2, \ldots, v_r\} \)
is called a chain based on \( v_1 \) if \( N^{k+1} v_k = v_k \) for all \( k = 1, 2, \ldots, r \) and \( v_1 \) is an eigenvector so
\( N v_1 = 0 \). It will be called a maximal chain if there is no solution \( v \) to the equation, \( N v = v_r \).
Now there will be a sequence of steps leading to the desired basis.

(a) Show that the vectors in any chain are linearly independent and for
\[
\{v_1, v_2, \ldots, v_r\}
\]
a chain based on \( v_1 \),
\[
N : \text{span} (v_1, v_2, \ldots, v_r) \mapsto \text{span} (v_1, v_2, \ldots, v_r).
\]  
(6.3)
Also if \( \{v_1, v_2, \ldots, v_r\} \) is a chain, then \( r \leq n \). Hint: If \( 0 = \sum_{i=1}^{r} c_i v_i \), and the last nonzero scalar occurs at \( t \), do \( N^{t-1} \) to the sum and see what happens to \( c_t \).

(b) Consider the set of all chains based on eigenvectors. Since all have total length no larger
than \( n \) it follows there exists one which has maximal length, \( \{v_1, \ldots, v_{r_1}\} \equiv B_1 \). If
\( \text{span} (B_1) \) contains all eigenvectors of \( N \), then stop. Otherwise, consider all chains based on
eigenvectors not in \( \text{span} (B_1) \) and pick one, \( B_2 \equiv \{v_{r_1}^2, \ldots, v_{r_2}^2\} \) which is as long as possible.
Thus \( r_2 \leq r_1 \). If \( \text{span} (B_1, B_2) \) contains all eigenvectors of \( N \), stop. Otherwise, consider
all chains based on eigenvectors not in \( \text{span} (B_1, B_2) \) and pick one, \( B_3 \equiv \{v_{r_2}^3, \ldots, v_{r_3}^3\} \)
such that \( r_3 \) is as large as possible. Continue this way. Thus \( r_k \geq r_{k+1} \). Then show that
the above process terminates with a finite list of chains
\[
\{B_1, \ldots, B_s\}
\]
because for any \( k, \{B_1, \ldots, B_k\} \) is linearly independent. Hint: From part a. you know
this is true if \( k = 1 \). Suppose true for \( k-1 \) and letting \( L (B_i) \) denote a linear combination
of vectors of \( B_i \), suppose
\[
\sum_{i=1}^{k} L (B_i) = 0
\]
Then we can assume \( L (B_k) \neq 0 \) by induction. Let \( v_1^k \) be the last term in \( L (B_k) \) which has
nonzero scalar. Now act on the whole thing with \( N^{k-1} \) to find \( v_1^k \) as a linear combination
of vectors in \( \{B_1, \ldots, B_{k-1}\} \), a contradiction to the construction. You fill in the details.
(c) Suppose \( Nw = 0 \) (\( w \) is an eigenvector). Show that there exist scalars, \( c_i \) such that

\[
w = \sum_{i=1}^{s} c_i v_i^i.
\]

Recall that \( v_i^i \) is the eigenvector in the \( i^{th} \) chain on which this chain is based. You know that \( w \) is a linear combination of the vectors in \( \{B_1, \ldots, B_s\} \). This says that in fact it is a linear combination of the bottom vectors in the \( B_i \). **Hint:** You know that

\[
w = \sum_{i=1}^{s} L(B_i).
\]

Let \( v_i^s \) be the last in \( L(B_s) \) which has nonzero scalar. Suppose that \( i > 1 \). Now do \( N^{i-1} \) to both sides and obtain that \( v_i^s \) is in the span of \( \{B_1, \ldots, B_{s-1}\} \) which is a contradiction. Hence \( i = 1 \) and so the only term of \( L(B_s) \) is one involving an eigenvector. Now do something similar to \( L(B_{s-1}), L(B_{s-2}) \) etc. You fill in details.

(d) If \( Nw = 0 \), then \( w \in \text{span}(B_1, \ldots, B_s) \). This was what was just shown. In fact, it was a particular linear combination involving the bases of the chains. What if \( N^k w = 0 \)? Does it still follow that \( w \in \text{span}(B_1, \ldots, B_s) \)? Show that if \( N^k w = 0 \), then \( w \in \text{span}(B_1, \ldots, B_s) \). **Hint:** Say \( k \) is as small as possible such that \( N^k w = 0 \). Then you have \( N^{k-1} w \) is an eigenvector and so

\[
N^{k-1} w = \sum_{i=1}^{s} c_i v_i^i
\]

If \( N^{k-1} w \) is the base of some chain \( B_i \), then there is nothing to show. Otherwise, consider the chain \( N^{k-1} w, N^{k-2} w, \ldots, w \). It cannot be any longer than any of the chains \( B_1, B_2, \ldots, B_s \) why? Therefore, \( v_1^i = N^{k-1} v_k^i \). Why is \( v_k^i \in B_i \)? This is where you use that this is no longer than any of the \( B_i \). Thus

\[
N^{k-1} \left( w - \sum_{i=1}^{s} c_i v_k^i \right) = 0
\]

By induction, (details) \( w - \sum_{i=1}^{s} c_i v_k^i \in \text{span}(B_1, \ldots, B_s) \).

(e) Since \( N \) is nilpotent, \( \ker(N^m) = V \) for some \( m \) and so all of \( V \) is in \( \text{span}(B_1, \ldots, B_s) \).

(f) Now explain why the matrix with respect to the ordered basis \( (B_1, \ldots, B_s) \) is the kind of thing desired and described in the above problem. Also explain why the size of the blocks decreases from upper left to lower right. To see why the matrix is like the above, consider

\[
\begin{pmatrix}
0 & v_1^i & \cdots & v_i^i & \cdots & v_{r_i-1}^i
\end{pmatrix} =
\begin{pmatrix}
v_1^i & v_2^i & \cdots & v_{r_i}^i
\end{pmatrix} M_i
\]

where \( M_i \) is the \( i^{th} \) block and \( r_i \) is the length of the \( i^{th} \) chain.

If you have gotten through this, then along with the previous problem, you have proved the existence of the Jordan canonical form, one of the greatest results in linear algebra. It will be considered a different way later. Specifically, you have shown that if the minimum polynomial splits, then the linear transformation has a matrix of the following form:

\[
J_{r_1}(\mu_1) \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & J_{r_s}(\mu_s) \end{pmatrix}
\]

where without loss of generality, you can arrange these blocks to be decreasing in size from the upper left to the lower right and \( J(\mu_i) \) is of the form

\[
J_{r_i}(\mu_i) \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & J_{r_i}(\mu_i) \end{pmatrix}
\]
Where $J_r(\mu_i)$ is the $r \times r$ matrix which is of the following form

$$J_r(\mu_i) = \begin{pmatrix} \mu_i & 1 & 0 \\ \ldots & \mu_i & \ldots \\ 0 & \ldots & \mu_i \end{pmatrix}$$

and the blocks $J_r(\mu_i)$ can also be arranged to have their size decreasing from the upper left to lower right.

12. (Extra important) The following theorem gives an easy condition for which the Jordan canonical form will be a diagonal matrix.

**Theorem 6.5.1** Let $A \in \mathcal{L}(V, V)$ and suppose $(u_i, \lambda_i), i = 1, 2, \ldots, m$ are eigen-pairs such that if $i \neq j$, then $\lambda_i \neq \lambda_j$. Then $\{u_1, \ldots, u_m\}$ is linearly independent. In words, eigenvectors from distinct eigenvalues are linearly independent.

**Hint:** Suppose $\sum_{i=1}^{k} c_i u_i = 0$ where $k$ is as small as possible such that not all of the $c_i = 0$. Then $c_k \neq 0$. Explain why $k > 1$ and

$$\sum_{i=1}^{k} c_i \lambda_k u_i = \sum_{i=1}^{k} c_i \lambda_i u_i$$

Now

$$\sum_{i=1}^{k} c_i (\lambda_k - \lambda_i) u_i = 0$$

Obtain a contradiction of some sort at this point. Thus if the $n \times n$ matrix has $n$ distinct eigenvalues, then the corresponding eigenvectors will be a linearly independent set and so the matrix will be diagonal and all the Jordan blocks will be single numbers.
Chapter 7

Canonical Forms

Linear algebra is really all about linear transformations and the fundamental question is whether a matrix comes from some linear transformation with respect to some basis. In other words, are two matrices really from the same linear transformation? As proved above, this happens if and only if the two are similar. Canonical forms allow one to answer this question. There are two main kinds of canonical form, the Jordan canonical form for the case where the minimum polynomial splits and the rational canonical form in the other case. Of the two, the Jordan canonical form is the one which is used the most in applied math. However, the other one is also pretty interesting.

7.1 Cyclic Sets

It was shown above that for $A \in \mathcal{L}(V, V)$ for $V$ a finite dimensional vector space over the field of scalars $F$, there exists a direct sum decomposition

$$V = V_1 \oplus \cdots \oplus V_q$$

where

$$V_k = \ker(\phi_k(A)^{m_k})$$

and $\phi_k(\lambda)$ is an irreducible monic polynomial. Here the minimum polynomial of $A$ was

$$\prod_{k=1}^q \phi_k(\lambda)^{m_k}$$

Next I will consider the problem of finding a basis for $V_k$ such that the matrix of $A$ restricted to $V_k$ assumes various forms.

**Definition 7.1.1** Letting $x \neq 0$ denote by $\beta_x$ the vectors $\{x, Ax, A^2x, \cdots, A^{m-1}x\}$ where $m$ is the smallest such that $A^m x \in \text{span}(x, \cdots, A^{m-1}x)$. This is called an $A$ cyclic set. The vectors which result are also called a Krylov sequence. For such a sequence of vectors, $|\beta_x| \equiv m$, the number of vectors in $\beta_x$.

Note that for such an $A$ cyclic set, there exists a unique monic polynomial $\eta(\lambda)$ of degree $|\beta_x|$ such that $\eta(A)x = 0$ such that if $\phi(A)x = 0$ for any other polynomial, then $\eta(\lambda)$ must divide $\phi(\lambda)$. If this is not clear, see below.

The first thing to notice is that such a Krylov sequence is always linearly independent.

**Lemma 7.1.2** Let $\beta_x = \{x, Ax, A^2x, \cdots, A^{m-1}x\}, x \neq 0$ where $m$ is the smallest such that $A^m x \in \text{span}(x, \cdots, A^{m-1}x)$. Then $\{x, Ax, A^2x, \cdots, A^{m-1}x\}$ is linearly independent.
The last claim follows because $A^mx$ is the span of vectors of the form $A^px$. ■

For more on the next lemma and the following theorem, see Hofman and Kunze [13]. I am following the presentation in Friedberg Insel and Spence [12]. See also Herstein [10] or Problem 11 on Page 114 for a different approach to canonical forms. To help organize the ideas in the lemma, here is a diagram.
Lemma 7.1.5 Let $W$ be an $A$ invariant ($AW \subseteq W$) subspace of $\ker(\phi(A)^m)$ for $m$ a positive integer where $\phi(\lambda)$ is an irreducible monic polynomial of degree $d$. Let $U$ be an $A$ invariant subspace of $\ker(\phi(A))$.

If $\{v_1, \ldots, v_s\}$ is a basis for $W$ then if $x \in U \setminus W$,

$$\{v_1, \ldots, v_s, \beta_x\}$$

is linearly independent. (In other words, we know that $\{v_1, \ldots, v_s, x\}$ is linearly independent by earlier theorems, but this says that you can include, not just $x$ but the entire $A$ cyclic set beginning with $x$.)

There exist vectors $x_1, \ldots, x_p$ each in $U$ such that

$$\{v_1, \ldots, v_s, \beta_{x_1}, \ldots, \beta_{x_p}\}$$

is a basis for $U + W$.

Also, if $x \in \ker(\phi(A)^m)$, $|\beta_x| = kd$ where $k \leq m$. Here $|\beta_x|$ is the length of $\beta_x$, the degree of the monic polynomial $\eta(\lambda)$ satisfying $\eta(A)x = 0$ with $\eta(\lambda)$ having smallest possible degree.

**Proof:** Claim: If $x \in \ker\phi(A)$, and $|\beta_x|$ denotes the length of $\beta_x$, then $|\beta_x| = d$ the degree of the irreducible polynomial $\phi(\lambda)$ and so

$$\beta_x = \{x, Ax, A^2x, \ldots, A^{d-1}x\}$$

also span $\beta_x$ is $A$ invariant, $A(\text{span}(\beta_x)) \subseteq \text{span}(\beta_x)$.

**Proof of the claim:** This was discussed earlier, but to repeat, let $m = |\beta_x|$. That is, there exists monic $\eta(\lambda)$ of degree $m$ and $\eta(A)x = 0$ with $m$ is as small as possible for this to happen. Then from the usual process of division of polynomials, there exist $l(\lambda), r(\lambda)$ such that $r(\lambda) = 0$ or else has smaller degree than that of $\eta(\lambda)$ such that

$$\phi(\lambda) = \eta(\lambda)l(\lambda) + r(\lambda)$$

If $\deg(r(\lambda)) < \deg(\eta(\lambda))$, then the equation implies $0 = \phi(A)x = r(A)x$ and so $m$ was incorrectly chosen. Hence $r(\lambda) = 0$ and so if $l(\lambda) \neq 1$, then $\eta(\lambda)$ divides $\phi(\lambda)$ contrary to the assumption that $\phi(\lambda)$ is irreducible. Hence $l(\lambda) = 1$ and $\eta(\lambda) = \phi(\lambda)$. The claim about span $\beta_x$ is obvious because $A^d x \in \text{span}(\beta_x)$. This shows the claim.

Suppose now $x \in U \setminus W$ where $U \subseteq \ker(\phi(A))$. Consider

$$\{v_1, \ldots, v_s, \beta_x\}.$$ 

Is this set of vectors independent? Suppose

$$\sum_{i=1}^{s} a_i v_i + \sum_{j=1}^{d} d_j A^{j-1} x = 0.$$
If \( z = \sum_{j=1}^{d} d_j A^{j-1} x \), then \( z \in W \cap \text{span} \{ x, A x, \cdots, A^{d-1} x \} \) because \( z = -\sum_{i=1}^{m} a_i v_i \). Then by Lemma \ref{lem:minimum-polynomial}, for each \( m \leq d - 1 \),

\[
A^m z \in W \cap \text{span} \{ x, A x, \cdots, A^{d-1} x \}
\]

Therefore,

\[
\text{span} \{ z, A z, \cdots, A^{d-1} z \} \subseteq W \cap \text{span} \{ x, A x, \cdots, A^{d-1} x \} \subseteq \text{span} \{ x, A x, \cdots, A^{d-1} x \}
\]  

Suppose \( z \neq 0 \). Then from the Lemma \ref{lem:minimum-polynomial} above, \( \{ z, A z, \cdots, A^{d-1} z \} \) must be linearly independent. Therefore,

\[
d = \dim \left( \text{span} \{ z, A z, \cdots, A^{d-1} z \} \right) \leq \dim \left( W \cap \text{span} \{ x, A x, \cdots, A^{d-1} x \} \right)
\]

\[
\leq \dim \left( \text{span} \{ x, A x, \cdots, A^{d-1} x \} \right) = d
\]

Thus

\[
W \cap \text{span} \{ x, A x, \cdots, A^{d-1} x \} = \text{span} \{ x, A x, \cdots, A^{d-1} x \}
\]

which would require \( W \supseteq \text{span} \{ x, A x, \cdots, A^{d-1} x \} \) so \( x \in W \) but this is assumed not to take place. Hence \( z = 0 \) and so the linear independence of the \( \{ v_1, \cdots, v_s \} \) implies each \( a_i = 0 \). Then the linear independence of \( \{ x, A x, \cdots, A^{d-1} x \} \), which follows from Lemma \ref{lem:minimum-polynomial}, shows each \( d_j = 0 \). Thus

\[
\{ v_1, \cdots, v_s, x, A x, \cdots, A^{d-1} x \}
\]

is linearly independent as claimed.

Let \( x \in U \setminus W \subseteq \ker (\phi(A)) \). Then it was just shown that \( \{ v_1, \cdots, v_s, \beta_x \} \) is linearly independent. Let \( W_1 \) be given by

\[
y \in \text{span} \{ v_1, \cdots, v_s, \beta_x \} \equiv W_1
\]

Then \( W_1 \) is \( A \) invariant. If \( W_1 \) equals \( U + W \), then you are done. If not, let \( W_1 \) play the role of \( W \) and pick \( x_1 \in U \setminus W_1 \) and repeat the argument. Continue till

\[
\text{span} \{ v_1, \cdots, v_s, \beta_{x_1}, \cdots, \beta_{x_n} \} = U + W
\]

The process stops because \( \ker (\phi(A)^m) \) is finite dimensional.

Finally, letting \( x \in \ker (\phi(A)^m) \), there is a monic polynomial \( \eta(\lambda) \) such that \( \eta(A) x = 0 \) and \( \eta(\lambda) \) is of smallest possible degree, which degree equals \( |\beta_x| \). Then

\[
\phi(\lambda)^m = \eta(\lambda) l(\lambda) + r(\lambda)
\]

If \( \deg(r(\lambda)) < \deg(\eta(\lambda)) \), then \( r(A)x = 0 \) and \( \eta(\lambda) \) was incorrectly chosen. Hence \( r(\lambda) = 0 \) and so \( \eta(\lambda) \) must divide \( \phi(\lambda)^m \). Hence by Corollary \ref{cor:division}, \( \eta(\lambda) = \phi(\lambda)^k \) where \( k \leq m \). Thus \( |\beta_x| = kd = \deg(\eta(\lambda)) \). 

With this preparation, here is the main result about a basis for \( V \) where \( A \in \mathcal{L}(V, V) \) and the minimum polynomial for \( A \) is \( \phi(\lambda)^m \) for \( \phi(\lambda) \) an irreducible monic polynomial. We have in mind \( V = V_i \) from \ref{lem:minimum-polynomial}. There is a very interesting generalization of this theorem in \ref{thm:canonical-forms} which pertains to the existence of complementary subspaces. For an outline of this generalization, see Problem \ref{prob:canonical-forms} on Page \ref{thm:canonical-forms}.

**Theorem 7.1.6** Suppose \( A \in \mathcal{L}(V, V) \) and the minimum polynomial of \( A \) is \( \phi(\lambda)^m \) where \( \phi(\lambda) \) is a monic irreducible polynomial. Then there exists a basis for \( V = \ker(\phi(A)^m) \) which is of the form \( \beta = \{ \beta_{x_1}, \cdots, \beta_{x_p} \} \).
Proof: First suppose \( m = 1 \). Then in Lemma 7.2.7, you can let \( W = \{0\} \) and \( U = \ker(\phi(A)) \). Then by this lemma, there exist \( v_1, v_2, \ldots, v_s \) such that \( \{\beta_{v_1}, \ldots, \beta_{v_s}\} \) is a basis for \( \ker(\phi(A)) \). Suppose then that the theorem is true for \( m - 1, m \geq 2 \).

Now the minimum polynomial for \( A \) on \( V \) is assumed to be \( \phi(A)^m \) where \( \phi(\lambda) \) is monic and irreducible. Then \( \phi(A)(V) \) is an invariant subspace of \( V \). What is the minimum polynomial of \( A \) on \( \phi(A)(V) \)? Clearly \( \phi(A)^{m-1} \) will send everything in \( \phi(A)(V) \) to 0. If \( \eta(\lambda) \) is the minimum polynomial of \( A \) on \( \phi(A)(V) \), then

\[
\phi(\lambda)^{m-1} = \eta(\lambda) + r(\lambda)
\]

and \( r(\lambda) \) must equal 0 since otherwise \( r(A) = 0 \) and \( \eta(\lambda) \) was not minimum. By Corollary 7.2.9, \( \eta(\lambda) = \phi(\lambda)^k \) for some \( k \leq m - 1 \). However, it cannot happen that \( k < m - 1 \) because if so, \( \phi(\lambda)^m \) would fail to be the minimum polynomial for \( A \) on \( V \). By induction, \( \phi(A)(V) \) has a basis

\[
\{\beta_{x_1}, \ldots, \beta_{x_p}\}.
\]

Let \( y_j \in V \) be such that \( \phi(A)y_j = x_j \). Consider \( \{\beta_{y_1}, \ldots, \beta_{y_p}\} \). Are these vectors independent? Suppose

\[
0 = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} A^{j-1} y_i \equiv \sum_{i=1}^{p} f_i(A) x_i \tag{7.3}
\]

If the sum involved \( x_i \) in place of \( y_i \), then something could be said because \( \{\beta_{x_1}, \ldots, \beta_{x_p}\} \) is a basis.

Do \( \phi(A) \) to both sides to obtain

\[
0 = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} A^{j-1} x_i \equiv \sum_{i=1}^{p} f_i(A) x_i
\]

Now \( f_i(A)x_i = 0 \) for each \( i \) since \( f_i(A)x_i \in \text{span}(\beta_{x_i}) \) and as just mentioned, \( \{\beta_{x_1}, \ldots, \beta_{x_p}\} \) is a basis. Let \( \eta_i(\lambda) \) be the monic polynomial of smallest degree such that \( \eta_i(A)x_i = 0 \). Then

\[
f_i(\lambda) = \eta_i(\lambda) l(\lambda) + r(\lambda)
\]

where \( r(\lambda) = 0 \) or else it has smaller degree than \( \eta_i(\lambda) \). However, the equation then shows that \( r(A)x_i = 0 \) which would contradict the choice of \( \eta_i(\lambda) \). Thus \( r(\lambda) = 0 \) and \( \eta_i(\lambda) \) divides \( f_i(\lambda) \). Also, \( \phi(\lambda)^{m-1}x_i = 0 \) and so \( \eta_i(\lambda) \) must divide \( \phi(\lambda)^{m-1} \). From Corollary 7.2.10, it follows that, since \( \phi(\lambda) \) is irreducible, \( \eta_i(\lambda) = \phi(\lambda)^k \) for some \( k \leq m - 1 \). Thus \( \phi(\lambda) \) divides \( \eta_i(\lambda) \) which divides \( f_i(\lambda) \). Hence \( f_i(\lambda) = \phi(\lambda) g_i(\lambda) \). Now

\[
0 = \sum_{i=1}^{p} f_i(A) y_i = \sum_{i=1}^{p} g_i(A) \phi(A) y_i = \sum_{i=1}^{p} g_i(A) x_i.
\]

By the same reasoning just given, since \( g_i(A)x_i \in \text{span}(\beta_{x_i}) \), it follows that each \( g_i(A)x_i = 0 \). Therefore, \( f_i(A)y_i = g_i(A)\phi(A)y_i = g_i(A)x_i = 0 \). Therefore,

\[
f_i(A)y_i = \sum_{j=1}^{p} a_{ij} A^{j-1} y_i = 0
\]

and by independence of the \( \beta_{y_j} \), this implies \( a_{ij} = 0 \).

Next, it follows from the definition that \( \phi(A) \text{span}(\beta_{y_k}) = \text{span}(\beta_{x_k}) \) and consequently, for \( W \equiv \text{span}(\beta_{y_1}, \ldots, \beta_{y_p}) \)

\[
\phi(A)(V) = \text{span}(\beta_{x_1}, \ldots, \beta_{x_p}) \subseteq \phi(A) \text{span}(\beta_{y_1}, \ldots, \beta_{y_p}) \equiv \phi(A)(W)
\]
To see the inclusion,
\[ A^r x_q = A^r \phi (A) y_q = \phi (A) A^r y_q \in \phi (A) \operatorname{span} \left( \beta_{y_q} \right) \subseteq \phi (A) \operatorname{span} \left( \beta_{y_1}, \cdots, \beta_{y_r} \right) \]
It follows from Lemma 7.1.6 that \( V = W + \ker (\phi (A)) \). From Lemma 7.1.5 this last subspace has a basis of the form \( \{ \beta_{y_1}, \cdots, \beta_{y_r}, \beta_{z_1}, \cdots, \beta_{z_k} \} \). 

In Lemma 7.1.5 we got a basis for \( W + U \) which was of the form \( \{ v_1, \cdots, v_s, \beta_{x_1}, \cdots, \beta_{x_p} \} \) where \( \{ v_1, \cdots, v_s \} \) was a basis of \( W \). Now this was obtained by picking a succession of \( x_i \) in \( U \) which were not in \( W \), each leading to \( \beta_{x_i} \) and \( |\beta_{x_i}| = d \). Therefore, if you had a basis \( \{ v_1, \cdots, v_s, \beta_{x_1}, \cdots, \beta_{x_p} \} \) for \( U + W \), you would need to have \( \sum_{i=1}^q |\beta_{x_i}| = \sum_{i=1}^p |\beta_{x_i}| \) and each \( |\beta_{x_i}| = d \). Therefore, you must have \( p = q \) and the lengths of each of the \( \beta_{x_i} \) are the same. Otherwise, you would have bases for \( U + W \) which have different lengths. Suppose you have a basis \( \{ \beta_{x_1}, \cdots, \beta_{x_p} \} \) for \( \ker (\phi (A)^m) \). If you look at the numbers \( |\beta_{x_i}| \), do you have a uniquely determined collection of these numbers?

This is true if \( m = 1 \) which was just explained. Indeed you get \( \ker (\phi (A)) = \operatorname{span} (\beta_{x_i} : |\beta_{x_i}| = d) \).

Now consider \( \ker (\phi (A)^2) \). It is the span of a collection of cycles also. If you have a cycle \( x \) contained in \( \ker (\phi (A)^2) \) where \( |\beta_{x_i}| > d \), then from Lemma 7.1.5, \( |\beta_{x_i}| = kd \) for some \( k > 1 \). Could \( k > 2 \)? If so, \( x \) would not be in \( \ker (\phi (A)^2) \) because the first exponent \( r \) for which \( A^r x \) is in the span of the \( A^k x \) for \( k < r \) would by definition be \( kd \) and so \( \phi (A)^2 x \) would fail to be 0. Therefore,

\[ \ker (\phi (A)^2) = \operatorname{span} (\beta_{x_i} : |\beta_{x_i}| = d) \oplus \operatorname{span} (\beta_{x_i} : |\beta_{x_i}| = 2d) \]

Continuing this way,

\[ \ker (\phi (A)^k) = \bigoplus_{k=1}^s \operatorname{span} (\beta_{x_i} : |\beta_{x_i}| = kd) \]

where the new symbol denotes a direct sum of the indicated subspaces. As pointed out, the number of \( \beta_{x_i} \) for which \( |\beta_{x_i}| = d \) is uniquely determined. Now \( \ker (\phi (A)^2) \) involves those cycles which have length \( 2d \) along with those which have length \( d \). Hence the number of cycles having length \( 2d \) is also uniquely determined depending on \( \ker (\phi (A)) \) and \( \ker (\phi (A)^2) \). Continuing this way, we find that the number of cycles of length \( kd \) is uniquely determined for each \( k = 1, 2, \cdots, m \). We state this as the following theorem.

**Theorem 7.1.7** Let \( \phi (\lambda) \) be irreducible and let \( A \in \mathcal{L} (V, V) \) where \( V \) is a vector space. Suppose \( \ker (\phi (A)^m) \) is finite dimensional. Then there are cycles \( \beta_{x_i} \) such that

\[ \ker (\phi (A)^m) = \operatorname{span} \left( \beta_{x_1}, \cdots, \beta_{x_p} \right) \]

and

\[ \ker (\phi (A)^m) = \bigoplus_{k=1}^m \operatorname{span} (\beta_{x_i} : |\beta_{x_i}| = kd) \]

where the number of cycles \( \beta_{x_i} \) which have length \( kd \) is uniquely determined. The new symbol denotes a direct sum of the indicated subspaces.

### 7.2 The Rational Canonical Form

Let \( V \) be a finite dimensional vector space and let \( A \in \mathcal{L} (V, V) \). Let the minimum polynomial be

\[ \prod_{i=1}^q \phi_i (\lambda)^{m_i} \]
where \( \phi_i(\lambda) \) is irreducible. Thus we can consider

\[
V_i = \ker (\phi_i(A)^{m_i})
\]

From Theorem 7.1.7,

\[
\ker (\phi_i(A)^{m_i}) = \text{span} \left( \beta_{x_1}, \ldots, \beta_{x_{m_i}} \right)
\]

where the number of cycles in the above having length \( kd \) is uniquely determined. Now consider what happens to the restriction of \( A \) to \( \text{span} \left( \beta_{x_1} \right) \) where \( |\beta_{x_1}| = kd_i \). What is its matrix with respect to the basis for \( \text{span} \left( \beta_{x_1} \right) \)? Using the usual gimmick for finding matrix

\[
\begin{pmatrix}
Ax_j & A^2x_j & \cdots & A^{kd_i-1}x_j & A^{kd_i}x_j
\end{pmatrix} = \begin{pmatrix}
x_j & Ax_j & \cdots & A^{kd_i-2}x_j & A^{kd_i-1}x_j
\end{pmatrix} M
\]

Now if \( \phi_i(\lambda)^k = \left( a_0 + a_1\lambda + \cdots + a_{kd_i-1}\lambda^{kd_i-1} + \lambda^{kd_i} \right) \), then

\[
A^{kd_i}x_j = - \left( a_0 + a_1A + \cdots + a_{kd_i-1}A^{kd_i-1} \right) x_j
\]

therefore, we can determine the matrix \( M \). It is of the form

\[
\begin{pmatrix}
0 & \cdots & \cdots & \cdots & -a_0 \\
1 & 0 & \cdots & \cdots & -a_1 \\
0 & 1 & \cdots & \cdots & -a_{kd_i-2} \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & \cdots & 1 & -a_{kd_i-1}
\end{pmatrix}
\]

This is called a companion matrix. Thus, with respect to the bases \( \left\{ \beta_{x_1}, \ldots, \beta_{x_{m_i}} \right\}_{i=1}^{q} \) the matrix of \( A \) is block diagonal with each block itself being block diagonal, the smaller blocks being of the form in *. Conversely, if you have such a matrix, it determines a collection of cyclic bases \( \left\{ \beta_{x_1}, \ldots, \beta_{x_{m_i}} \right\}_{i=1}^{q} \). Corresponding to one of the blocks above, you would have

\[
\begin{pmatrix}
Az_1 & Az_2 & \cdots & Az_{kd_i-2} & Az_{kd_i-1}
\end{pmatrix} = \begin{pmatrix}
z_1 & z_2 & \cdots & z_{kd_i-2} & z_{kd_i-1}
\end{pmatrix} M
\]

and if \( M \) is of the form given above, you would need \( z_2 = Az_1, z_3 = Az_2 \) etc. This yields just such a cyclic basis with the last entry on the left \( A^{kd_i}z_1 \) which will then be equal to the appropriate linear combination of lower powers times \( z_1 \). Thus if the blocks corresponding to \( \ker (\phi_i(A)^{m_i}) \) are ordered to decrease in size from upper left to lower right, the matrix obtained is uniquely determined. From Theorem 7.1.7 the numbers of sub blocks of a given size in the block for \( \ker (\phi_i(A)^{m_i}) \) are uniquely determined. This matrix is called the rational canonical form. Thus the rational canonical form exists and is unique because if you had two different ones resulting from the same \( A \) then there would be a discrepancy in the numbers of cycles of a particular length.

**Theorem 7.2.1** Let \( A \in \mathcal{L}(V, V) \) where \( V \) is a finite dimensional vector space and let its minimum polynomial be

\[
\prod_{i=1}^{q} \phi_i(\lambda)^{m_i}
\]
Then there is a block diagonal matrix
\[
\begin{pmatrix}
M_1 & 0 \\
& \ddots \\
0 & M_q
\end{pmatrix}
\]
where \( M_i \) is the matrix of \( A \) with respect to \( \ker(\phi_i(\lambda)^{m_i}) \) taken with respect to a cyclic basis of the form \( \{\beta_{x_1}, \ldots, \beta_{x_p}\} \) and it is of the form
\[
\begin{pmatrix}
C_1 & 0 \\
& \ddots \\
0 & C_i
\end{pmatrix}
\]
where each \( C_j \) is a companion matrix as described in *. If we arrange these \( C_j \) to be descending in size from upper left to lower right, then the matrix just described is uniquely determined because the size of the \( C_i \) is determined by the numbers \( |\beta_{x_i}| \).

Sometimes we describe the uniqueness assertion by stating that the rational canonical form is unique up to order of blocks. Note that there are exactly two things which determine the rational canonical form, the factorization of the minimum polynomial into irreducible factors and the numbers consisting of \( |\beta_{x_i}| \) for \( \beta_{x_i} \) a cycle in a basis of \( V \). Thus, if you can find these two things, you can obtain the rational canonical form.

The important thing about this canonical form is that it does not depend on being able to factor the minimum polynomial into linear factors. However, in most cases of interest when the field of scalars is the complex numbers, one can factor the minimum polynomial. When this happens, there is a much better canonical form called the Jordan canonical form. We lead up to this by considering the special case of nilpotent transformations. This kind of thing occurs when in one of these cycles, you have \( A^m x = 0 \), not just that it is a linear combination of \( A^r x \) for \( r < m \). When we understand this special case, then the Jordan form becomes easy.

### 7.3 Nilpotent Transformations

**Definition 7.3.1** Let \( V \) be a vector space over the field of scalars \( \mathbb{F} \). Then \( N \in \mathcal{L}(V, V) \) is called nilpotent if for some \( m \), it follows that \( N^m = 0 \).

The following lemma contains some significant observations about nilpotent transformations.

**Lemma 7.3.2** Suppose \( N^k x \neq 0 \). Then \( \{x, N x, \ldots, N^k x\} \) is linearly independent. Also, the minimum polynomial of \( N \) is \( \lambda^m \) where \( m \) is the first such that \( N^m = 0 \).

**Proof:** Suppose \( \sum_{i=0}^{k} c_i N^i x = 0 \) where not all \( c_i = 0 \). There exists \( l \) such that \( k \leq l < m \) and \( N^{l+1} x = 0 \) but \( N^l x \neq 0 \). Then multiply both sides by \( N^l \) to conclude that \( c_0 = 0 \). Next multiply both sides by \( N^{l-1} \) to conclude that \( c_1 = 0 \) and continue this way to obtain that all the \( c_i = 0 \).

Next consider the claim that \( \lambda^m \) is the minimum polynomial. If \( p(\lambda) \) is the minimum polynomial, then by the division algorithm,
\[
\lambda^m = p(\lambda) l(\lambda) + r(\lambda)
\]
where the degree of \( r(\lambda) \) is less than that of \( p(\lambda) \) or else \( r(\lambda) = 0 \). The above implies \( 0 = 0 + r(N) \) contrary to \( p(\lambda) \) being minimum. Hence \( r(\lambda) = 0 \) and so \( p(\lambda) \) divides \( \lambda^m \). Hence \( p(\lambda) = \lambda^k \) for \( k \leq m \). But if \( k < m \), this would contradict the definition of \( m \) as being the smallest such that \( N^m = 0 \).
For such a nilpotent transformation, let \( \{ \beta_{x_1}, \cdots, \beta_{x_q} \} \) be a basis for \( \ker (N^m) = V \) where these \( \beta_{x_i} \) are cyclic. This basis exists thanks to Theorem 7.1.6. Note that you can have \( |\beta_x| < m \) because it is possible for \( N^k x = 0 \) without \( N^k = 0 \). Thus

\[
V = \text{span} \left( \beta_{x_1} \right) \oplus \cdots \oplus \text{span} \left( \beta_{x_q} \right),
\]
each of these subspaces in the above direct sum being \( N \) invariant. For \( x \) one of the \( x_k \), consider \( \beta_x \) given by

\[
x, N x, N^2 x, \cdots, N^{r_k - 1} x
\]
where \( N^r x \) is in the span of the above vectors. Thus it must be 0 since if not, you could add it to the list and by the above lemma, the longer list would be linearly independent contrary to \( \beta_x \) being equal to the above list.

By Theorem 6.0.3, the matrix of \( N \) with respect to the above basis is the block diagonal matrix

\[
\begin{pmatrix}
M^1 & 0 \\
& \ddots \\
0 & \cdots & M^q
\end{pmatrix}
\]
where \( M^k \) denotes the matrix of \( N \) restricted to \( \text{span} \left( \beta_{x_k} \right) \). In computing this matrix, I will order \( \beta_{x_k} \) as follows:

\[
(N^{r_k - 1} x_k, \cdots, x_k)
\]
Also the cyclic sets \( \beta_{x_1}, \beta_{x_2}, \cdots, \beta_{x_q} \) will be ordered according to length, the length of \( \beta_{x_i} \) being at least as large as the length of \( \beta_{x_{i+1}} \), \( |\beta_{x_k}| \equiv r_k \). This is done to make the ones come out on the top of the main diagonal rather than below it as was the case with the rational canonical form mentioned above.

Why do we do this? I am not sure, but it seems to be traditional. I think it may result from the methods used in ordinary differential equations to find solutions to first order systems when you look for generalized eigenvectors. Then since \( N^{r_k} x_k = 0 \), it is now easy to find \( M^k \). Using the procedure mentioned above for determining the matrix of a linear transformation,

\[
\begin{pmatrix}
0 & N^{r_k - 1} x_k & \cdots & N x_k \\
N^{r_k - 1} x_k & N^{r_k - 2} x_k & \cdots & x_k
\end{pmatrix}
\]
Thus the matrix \( M_k \) is the \( r_k \times r_k \) matrix which has ones down the super diagonal and zeros elsewhere. Note that the size of the blocks is determined by \( |\beta_{x_i}| \) just as in the above rational canonical form. Since the numbers of these cycles having length \( k \) is uniquely determined, the matrix for \( N \) is also uniquely determined up to order of the blocks. The following convenient notation will be used.

**Definition 7.3.3** \( J_k (\alpha) \) is a Jordan block if it is a \( k \times k \) matrix of the form

\[
J_k (\alpha) = \begin{pmatrix}
\alpha & 1 & 0 \\
0 & \ddots & \ddots \\
\vdots & \ddots & \ddots & 1 \\
0 & \cdots & 0 & \alpha
\end{pmatrix}
\]
In words, there is an unbroken string of ones down the super diagonal and the number \( \alpha \) filling every space on the main diagonal with zeros everywhere else.
CHAPTER 7. CANONICAL FORMS

Then with this definition and the above discussion, the following proposition has been proved.

**Proposition 7.3.4** Let $N \in \mathcal{L}(W, W)$ be nilpotent,

$$N^m = 0$$

for some $m \in \mathbb{N}$. Here $W$ is a $p$ dimensional vector space with field of scalars $\mathbb{F}$. Then there exists a basis for $W$ such that the matrix of $N$ with respect to this basis is of the form

$$J = \begin{pmatrix} J_{r_1}(0) & 0 \\ J_{r_2}(0) & \ddots \\ \vdots & \ddots & \ddots \\ 0 & \cdots & 0 & J_{r_s}(0) \end{pmatrix}$$

(7.4)

where $r_1 \geq r_2 \geq \cdots \geq r_s \geq 1$ and $\sum_{i=1}^{s} r_i = p$. In the above, the $J_{r_j}(0)$ is called a Jordan block of size $r_j \times r_j$ with 0 down the main diagonal.

**Observation 7.3.5** Observe that $J_{r}(0)^{r} = 0$ but $J_{r}(0)^{r-1} \neq 0$.

In fact, the matrix of the above proposition is unique. This was proved in Theorem 7.2.1 but here we give a simpler explanation.

**Corollary 7.3.6** Let $J, J'$ both be matrices of the nilpotent linear transformation $N \in \mathcal{L}(W, W)$ which are of the form described in Proposition 7.3.4. Then $J = J'$. In fact, if the rank of $J^{r}$ equals the rank of $J'^{r}$ for all nonnegative integers $k$, then $J = J'$.

**Proof:** Since $J$ and $J'$ are similar, it follows that for each $k$ an integer, $J^k$ and $J'^k$ are similar. Hence, for each $k$, these matrices have the same rank. Now suppose $J \neq J'$. Note first that

$$J_r(0)^r = 0, J_r(0)^{r-1} \neq 0.$$ 

Denote the blocks of $J$ as $J_{r_k}(0)$ and the blocks of $J'$ as $J'_{r_k}(0)$. Let $k$ be the first such that $J_{r_k}(0) \neq J'_{r_k}(0)$. Suppose that $r_k > r'_k$. By block multiplication and the above observation, it follows that the two matrices $J^{r_k-1}$ and $J'^{r_k-1}$ are respectively of the forms

$$\begin{pmatrix} M_{r_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & M_{r_k} \end{pmatrix}, \begin{pmatrix} M'_{r_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & M'_{r_k} \end{pmatrix}$$

where $M_{r_j} = M'_{r_j}$ for $j \leq k - 1$ but $M_{r_k}$ is a zero $r'_k \times r'_k$ matrix while $M_{r_k}$ is a larger matrix which is not equal to 0. For example, $M_{r_k}$ could look like

$$M_{r_k} = \begin{pmatrix} 0 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 0 & 0 \end{pmatrix}$$

Thus there are more pivot columns in $J^{r_k-1}$ than in $(J')^{r_k-1}$, contradicting the requirement that $J^k$ and $J'^k$ have the same rank. ■
7.4 The Jordan Canonical Form

The Jordan canonical form has to do with the case where the minimum polynomial of \( A \in \mathcal{L}(V, V) \) splits, meaning that it can be completely factored. Thus there exist \( \lambda_k \) in the field of scalars such that the minimum polynomial of \( A \) is of the form

\[
p(\lambda) = \prod_{k=1}^{r} (\lambda - \lambda_k)^{m_k}
\]

Recall the following which follows from Theorem 6.3.2.

**Proposition 7.4.1** Let the minimum polynomial of \( A \in \mathcal{L}(V, V) \) be given by

\[
p(\lambda) = \prod_{k=1}^{r} (\lambda - \lambda_k)^{m_k}
\]

Then the eigenvalues of \( A \) are \( \{\lambda_1, \cdots, \lambda_r\} \).

It follows from Theorem 6.1.10 that

\[
V = \ker (A - \lambda_1 I)^{m_1} \oplus \cdots \oplus \ker (A - \lambda_r I)^{m_r}
\]

\[
\equiv V_1 \oplus \cdots \oplus V_r
\]

where \( I \) denotes the identity linear transformation. Without loss of generality, let the dimensions of the \( V_k \) be decreasing as \( k \) increases. These \( V_k \) are called the generalized eigenspaces.

It follows from the definition of \( V_k \) that \( (A - \lambda_k I) \) is nilpotent on \( V_k \) and clearly each \( V_k \) is \( A \)-invariant. Therefore from Proposition 7.3.4 and letting \( A_k \) denote the restriction of \( A \) to \( V_k \), there exists an ordered basis for \( V_k, \beta_k \) such that with respect to this basis, the matrix of \( (A_k - \lambda_k I) \) is of the form given in that proposition, denoted here by \( J^k \). What is the matrix of \( A_k \) with respect to \( \beta_k^k \) Letting \( \{b_1, \cdots, b_r\} = \beta_k^k \),

\[
A_k b_j = (A_k - \lambda_k I) b_j + \lambda_k I b_j = \sum s J^k_{k} b_s + \sum s \lambda_k \delta_{s j} b_s = \sum s (J^k_{k} + \lambda_k \delta_{s j}) b_s
\]

and so the matrix of \( A_k \) with respect to this basis is \( J^k + \lambda_k I \) where \( I \) is the identity matrix.

Therefore, with respect to the ordered basis \( \{\beta_1, \cdots, \beta_r\} \) the matrix of \( A \) is in Jordan canonical form. This means the matrix is of the form

\[
\begin{pmatrix}
J(\lambda_1) & 0 & & \\
& \ddots & & \\
& & J(\lambda_r) & \\
& & & 0
\end{pmatrix}
\]

(7.5)

where \( J(\lambda_k) \) is an \( m_k \times m_k \) matrix of the form

\[
\begin{pmatrix}
J_{k_1}(\lambda_k) & 0 & & \\
& J_{k_2}(\lambda_k) & & \\
& & \ddots & & \\
& & & J_{k_r}(\lambda_k)
\end{pmatrix}
\]

(7.6)

where \( k_1 \geq k_2 \geq \cdots \geq k_r \geq 1 \) and \( \sum_{i=1}^{r} k_i = m_k \). Here \( J_k(\lambda) \) is a \( k \times k \) Jordan block of the form

\[
\begin{pmatrix}
\lambda & 1 & 0 & & \\
0 & \lambda & \ddots & & \\
& \ddots & \ddots & \ddots & \\
0 & & & 0 & \lambda
\end{pmatrix}
\]

(7.7)
This proves the existence part of the following fundamental theorem.

Note that if any of the $k_i$ consists of eigenvectors, then the corresponding Jordan block will consist of a diagonal matrix having $\lambda_k$ down the main diagonal. This corresponds to $m_k = 1$. The vectors which are in $\ker (A - \lambda_k I)^{m_k}$ which are not in $\ker (A - \lambda_k I)$ are called generalized eigenvectors.

The following is the main result on the Jordan canonical form.

**Theorem 7.4.2** Let $V$ be an $n$ dimensional vector space with field of scalars $\mathbb{C}$ or some other field such that the minimum polynomial of $A \in \mathcal{L}(V, V)$ completely factors into powers of linear factors. Then there exists a unique Jordan canonical form for $A$ as described in 7.5 - 7.7, where uniqueness is in the sense that any two have the same number and size of Jordan blocks.

**Proof:** It only remains to verify uniqueness. Suppose there are two, $J$ and $J'$. Then these are matrices of $A$ with respect to possibly different bases and so they are similar. Therefore, they have the same minimum polynomials and the generalized eigenspaces have the same dimension. Thus the size of the matrices $J(\lambda_k)$ and $J'(\lambda_k)$ defined by the dimension of these generalized eigenspaces, also corresponding to the algebraic multiplicity of $\lambda_k$, must be the same. Therefore, they comprise the same set of positive integers. Thus listing the eigenvalues in the same order, corresponding blocks $J(\lambda_k), J'(\lambda_k)$ are the same size.

It remains to show that $J(\lambda_k)$ and $J'(\lambda_k)$ are not just the same size but also are the same up to order of the Jordan blocks running down their respective diagonals. It is only necessary to worry about the number and size of the Jordan blocks making up $J(\lambda_k)$ and $J'(\lambda_k)$. Since $J, J'$ are similar, so are $J - \lambda_k I$ and $J' - \lambda_k I$.

Thus the following two matrices are similar

\[
A = \begin{pmatrix}
J(\lambda_1) - \lambda_k I & 0 \\
\vdots & \ddots & \ddots \\
0 & \cdots & J(\lambda_r) - \lambda_k I
\end{pmatrix}
\]

\[
B = \begin{pmatrix}
J'(\lambda_1) - \lambda_k I & 0 \\
\vdots & \ddots & \ddots \\
0 & \cdots & J'(\lambda_r) - \lambda_k I
\end{pmatrix}
\]

and consequently, $\text{rank } (A^k) = \text{rank } (B^k)$ for all $k \in \mathbb{N}$. Also, both $J(\lambda_j) - \lambda_k I$ and $J'(\lambda_j) - \lambda_k I$ are one to one for every $\lambda_j \neq \lambda_k$. Since all the blocks in both of these matrices are one to one except the blocks $J'(\lambda_k) - \lambda_k I, J(\lambda_k) - \lambda_k I$, it follows that this requires the two sequences of numbers $\{\text{rank } ((J(\lambda_k) - \lambda_k I)^m)\}_{m=1}^\infty$ and $\{\text{rank } ((J'(\lambda_k) - \lambda_k I)^m)\}_{m=1}^\infty$ must be the same.

Then

\[
J(\lambda_k) - \lambda_k I \equiv \begin{pmatrix}
J_{k_1}(0) & 0 \\
J_{k_2}(0) & \ddots \\
0 & \cdots & J_{k_r}(0)
\end{pmatrix}
\]
and a similar formula holds for $J'(\lambda_k)$

$$J'(\lambda_k) - \lambda_k I = \begin{pmatrix} J_{l_1}(0) & 0 \\ J_{l_2}(0) & \ddots \\ 0 & \cdots & J_{l_p}(0) \end{pmatrix}$$

and it is required to verify that $p = r$ and that the same blocks occur in both. Without loss of generality, let the blocks be arranged according to size with the largest on upper left corner falling to smallest in lower right. Now the desired conclusion follows from Corollary 7.3.6.

Note that if any of the generalized eigenspaces $\ker (A - \lambda_k I)^m$ has a basis of eigenvectors, then it would be possible to use this basis and obtain a diagonal matrix in the block corresponding to $\lambda_k$. By uniqueness, this is the block corresponding to the eigenvalue $\lambda_k$. Thus when this happens, the block in the Jordan canonical form corresponding to $\lambda_k$ is just the diagonal matrix having $\lambda_k$ down the diagonal and there are no generalized eigenvectors.

The Jordan canonical form is very significant when you try to understand powers of a matrix. There exists an $n \times n$ matrix $S$ such that

$$A = S^{-1}JS.$$ 

Therefore, $A^2 = S^{-1}JSS^{-1}JS = S^{-1}J^2S$ and continuing this way, it follows

$$A^k = S^{-1}J^kS.$$ 

where $J$ is given in the above corollary. Consider $J^k$. By block multiplication,

$$J^k = \begin{pmatrix} J_1^k & 0 \\ \vdots & \ddots \\ 0 & \cdots & J_p^k \end{pmatrix}.$$ 

The matrix $J_s$ is an $m_s \times m_s$ matrix which is of the form

$$J_s = D + N \quad (7.8)$$

for $D$ a multiple of the identity and $N$ an upper triangular matrix with zeros down the main diagonal. Thus $N^{m_s} = 0$. Now since $D$ is just a multiple of the identity, it follows that $DN = ND$. Therefore, the usual binomial theorem may be applied and this yields the following equations for $k \geq m_s$.

$$J_s^k = (D + N)^k = \sum_{j=0}^{k} \binom{k}{j} D^{k-j} N^j = \sum_{j=0}^{m_s} \binom{k}{j} D^{k-j} N^j, \quad (7.9)$$

the third equation holding because $N^{m_s} = 0$. Thus $J_s^k$ is of the form

$$J_s^k = \begin{pmatrix} \alpha^k & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha^k \end{pmatrix}.$$ 

Lemma 7.4.3 Suppose $J$ is of the form $J_s$ described above in (7.8) where the constant $\alpha$, on the main diagonal is less than one in absolute value. Then

$$\lim_{k \to \infty} (J_s^k)_{ij} = 0.$$ 

\footnote{The $S$ here is written as $S^{-1}$ in the corollary.}
Proof: From [7.9], it follows that for large \(k\), and \(j \leq m_s\),
\[
\binom{k}{j} \leq \frac{k(k-1) \cdots (k-m_s+1)}{m_s!}.
\]
Therefore, letting \(C\) be the largest value of \(|(N^{j})_{pq}|\) for \(0 \leq j \leq m_s\),
\[
|(J^{k})_{pq}| \leq m_s C \left( \frac{k(k-1) \cdots (k-m_s+1)}{m_s!} \right) |\alpha|^{k-m_s}
\]
which converges to zero as \(k \to \infty\). This is most easily seen by applying the ratio test to the series
\[
\sum_{k=m_s}^{\infty} \left( \frac{k(k-1) \cdots (k-m_s+1)}{m_s!} \right) |\alpha|^{k-m_s}
\]
and then noting that if a series converges, then the \(k^{th}\) term converges to zero.

7.5 Exercises

1. In the discussion of Nilpotent transformations, it was asserted that if two \(n \times n\) matrices \(A, B\) are similar, then \(A^k\) is also similar to \(B^k\). Why is this so? If two matrices are similar, why must they have the same rank?

2. If \(A, B\) are both invertible, then they are both row equivalent to the identity matrix. Are they necessarily similar? Explain.

3. Suppose you have two nilpotent matrices \(A, B\) and \(A^k\) and \(B^k\) both have the same rank for all \(k \geq 1\). Does it follow that \(A, B\) are similar? What if it is not known that \(A, B\) are nilpotent? Does it follow then?

4. (Review problem.) When we say a polynomial equals zero, we mean that all the coefficients equal 0. If we assign a different meaning to it which says that a polynomial \(p(\lambda)\) equals zero when it is the zero function, \((p(\lambda) = 0\) for every \(\lambda \in \mathbb{F}\)) does this amount to the same thing? Is there any difference in the two definitions for ordinary fields like \(\mathbb{Q}\)? Hint: Consider for the field of scalars \(\mathbb{Z}_2\), the integers mod 2 and consider \(p(\lambda) = \lambda^2 + \lambda\).

5. Let \(A \in \mathcal{L}(V, V)\) where \(V\) is a finite dimensional vector space with field of scalars \(\mathbb{F}\). Let \(p(\lambda)\) be the minimum polynomial and suppose \(\phi(\lambda)\) is any nonzero polynomial such that \(\phi(A)\) is not one to one and \(\phi(\lambda)\) has smallest possible degree such that \(\phi(A)\) is nonzero and not one to one. Show \(\phi(\lambda)\) must divide \(p(\lambda)\).

6. Let \(A \in \mathcal{L}(V, V)\) where \(V\) is a finite dimensional vector space with field of scalars \(\mathbb{F}\). Let \(p(\lambda)\) be the minimum polynomial and suppose \(\phi(\lambda)\) is an irreducible polynomial with the property that \(\phi(A) x = 0\) for some specific \(x \neq 0\). Show that \(\phi(\lambda)\) must divide \(p(\lambda)\). Hint: First write \(p(\lambda) = \phi(\lambda) g(\lambda) + r(\lambda)\) where \(r(\lambda)\) is either 0 or has degree smaller than the degree of \(\phi(\lambda)\). If \(r(\lambda) = 0\) you are done. Suppose it is not 0. Let \(\eta(\lambda)\) be the monic polynomial of smallest degree with the property that \(\eta(A) x = 0\). Now use the Euclidean algorithm to divide \(\phi(\lambda)\) by \(\eta(\lambda)\). Contradict the irreducibility of \(\phi(\lambda)\).

7. Let
\[
A = \begin{pmatrix}
1 & 0 & 0 \\
0 & 0 & -1 \\
0 & 1 & 0
\end{pmatrix}
\]
Find the minimum polynomial for \(A\).
8. Suppose \( A \) is an \( n \times n \) matrix and let \( \mathbf{v} \) be a vector. Consider the \( A \) cyclic set of vectors \( \{ \mathbf{v}, A\mathbf{v}, \ldots, A^{m-1}\mathbf{v} \} \) where this is an independent set of vectors but \( A^m\mathbf{v} \) is a linear combination of the preceding vectors in the list. Show how to obtain a monic polynomial of smallest degree, \( m, \phi_\mathbf{v}(\lambda) \) such that

\[
\phi_\mathbf{v}(A)\mathbf{v} = \mathbf{0}
\]

Now let \( \{ \mathbf{w}_1, \ldots, \mathbf{w}_n \} \) be a basis and let \( \phi(\lambda) \) be the least common multiple of the \( \phi_{\mathbf{w}_k}(\lambda) \). Explain why this must be the minimum polynomial of \( A \). Give a reasonably easy algorithm for computing \( \phi_\mathbf{v}(\lambda) \).

9. Here is a matrix.

\[
\begin{pmatrix}
-7 & -1 & -1 \\
-21 & -3 & -3 \\
70 & 10 & 10
\end{pmatrix}
\]

Using the process of Problem 8 find the minimum polynomial of this matrix. Determine whether it can be diagonalized from its minimum polynomial.

10. Let \( A \) be an \( n \times n \) matrix with field of scalars \( \mathbb{C} \) or more generally, the minimum polynomial splits. Letting \( \lambda \) be an eigenvalue, show the dimension of the eigenspace equals the number of Jordan blocks in the Jordan canonical form which are associated with \( \lambda \). Recall the eigenspace is \( \ker (\lambda I - A) \).

11. For any \( n \times n \) matrix, why is the dimension of the eigenspace always less than or equal to the algebraic multiplicity of the eigenvalue as a root of the characteristic equation? **Hint:** Note the algebraic multiplicity is the size of the appropriate block in the Jordan form.

12. Give an example of two nilpotent matrices which are not similar but have the same minimum polynomial if possible.

13. Here is a matrix. Find its Jordan canonical form by directly finding the eigenvectors and generalized eigenvectors based on these to find a basis which will yield the Jordan form. The eigenvalues are 1 and 2.

\[
\begin{pmatrix}
-3 & -2 & 5 & 3 \\
-1 & 0 & 1 & 2 \\
-4 & -3 & 6 & 4 \\
-1 & -1 & 1 & 3
\end{pmatrix}
\]

Why is it typically impossible to find the Jordan canonical form?

14. Let \( A \) be an \( n \times n \) matrix and let \( J \) be its Jordan canonical form. Here \( \mathbb{F} = \mathbb{R} \) or \( \mathbb{C} \). Recall \( J \) is a block diagonal matrix having blocks \( J_k(\lambda) \) down the diagonal. Each of these blocks is of the form

\[
J_k(\lambda) = \begin{pmatrix}
\lambda & 1 \\
& \ddots \\
& & \ddots \\
0 & & 1 & \lambda
\end{pmatrix}
\]

Now for \( \varepsilon > 0 \) given, let the diagonal matrix \( D_\varepsilon \) be given by

\[
D_\varepsilon = \begin{pmatrix}
1 & 0 \\
& \varepsilon \\
& & \ddots \\
0 & & & \varepsilon^{k-1}
\end{pmatrix}
\]
Show that $D^{-1}J_k(\lambda)D$ has the same form as $J_k(\lambda)$ but instead of ones down the super diagonal, there is $\varepsilon$ down the super diagonal. That is $J_k(\lambda)$ is replaced with
\[
\begin{pmatrix}
\lambda & \varepsilon & 0 \\
& \lambda & \ddots \\
& & \ddots & \varepsilon \\
0 & & & \lambda
\end{pmatrix}
\]

Now show that for $A$ an $n \times n$ matrix, it is similar to one which is just like the Jordan canonical form except instead of the blocks having 1 down the super diagonal, it has $\varepsilon$.

15. Let $A$ be in $\mathcal{L}(V,V)$ and suppose that $A^px \neq 0$ for some $x \neq 0$. Show that $A^p e_k \neq 0$ for some $e_k \in \{e_1, \cdots, e_n\}$, a basis for $V$. If you have a matrix which is nilpotent, $(A^m = 0$ for some $m$) will it always be possible to find its Jordan form? Describe how to do it if this is the case. **Hint:** First explain why all the eigenvalues are 0. Then consider the way the Jordan form for nilpotent transformations was constructed in the above.

16. Show that if two $n \times n$ matrices $A, B$ are similar, then they have the same minimum polynomial and also that if this minimum polynomial is of the form $p(\lambda) = \prod_{i=1}^{s} \phi_i(\lambda)^{r_i}$ where the $\phi_i(\lambda)$ are irreducible and monic, then $\ker(\phi_i(A)^{r_i})$ and $\ker(\phi_i(B)^{r_i})$ have the same dimension. Why is this so? This was what was responsible for the blocks corresponding to an eigenvalue being of the same size.

17. In Theorem 7.1.6 show that each cyclic set $\beta_x$ is associated with a monic polynomial $\eta_x(\lambda)$ such that $\eta_x(A)(x) = 0$ and this polynomial has smallest possible degree such that this happens. Show that the cyclic sets $\beta_x$ can be arranged such that $\eta_{x_1+1}(\lambda)/\eta_{x_1}(\lambda)$.

18. Show that if $A$ is a complex $n \times n$ matrix, then $A$ and $A^T$ are similar. **Hint:** Consider a Jordan block. Note that
\[
\begin{pmatrix}
0 & 0 & 1 \\
0 & 1 & 0 \\
1 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
\lambda & 1 & 0 \\
0 & \lambda & 1 \\
0 & 0 & \lambda
\end{pmatrix}
\begin{pmatrix}
0 & 0 & 1 \\
0 & 1 & 0 \\
1 & 0 & 0
\end{pmatrix}
= \begin{pmatrix}
\lambda & 0 & 0 \\
1 & \lambda & 0 \\
0 & 1 & \lambda
\end{pmatrix}
\]

19. (Extra important) Let $A$ be an $n \times n$ matrix. The trace of $A$, $\text{trace}(A)$ is defined as $\sum_i A_{ii}$. It is just the sum of the entries on the main diagonal. Show $\text{trace}(A) = \text{trace}(A^T)$. Suppose $A$ is $m \times n$ and $B$ is $n \times m$. Show that $\text{trace}(AB) = \text{trace}(BA)$. Now show that if $A$ and $B$ are similar $n \times n$ matrices, then $\text{trace}(A) = \text{trace}(B)$. Recall that $A$ is similar to $B$ means $A = S^{-1}BS$ for some matrix $S$.

20. (Extra important) If $A$ is an $n \times n$ matrix and the minimum polynomial splits in $\mathbb{F}$ the field of scalars, show that trace $(A)$ equals the sum of the eigenvalues listed according to multiplicity according to number of times they occur on the Jordan form. Next, show that this is true even if the minimum polynomial does not split.

21. Let $A$ be a linear transformation defined on a finite dimensional vector space $V$. Let the minimum polynomial be $\prod_{i=1}^{s} \phi_i(\lambda)^{m_i}$ and let $\{\beta_i^{v_1}, \cdots, \beta_i^{v_{l_i}}\}$ be the cyclic sets such that $\{\beta_i^{v_1}, \cdots, \beta_i^{v_{l_i}}\}$ is a basis for $\ker(\phi_i(A)^{m_i})$. Let $v = \sum_i \sum_j v_j$. Now let $q(\lambda)$ be any polynomial and suppose that
\[
q(A)v = 0
\]
Show that it follows $q(A) = 0$. **Hint:** First consider the special case where a basis for $V$ is $\{x, Ax, \cdots, A^{n-1}x\}$ and $q(A)x = 0$. 

7.6 Companion Matrices

Recall the concept of a companion matrix.

**Definition 7.6.1** Let
\[ q(\lambda) = a_0 + a_1\lambda + \cdots + a_{n-1}\lambda^{n-1} + \lambda^n \]
be a monic polynomial. The companion matrix of \( q(\lambda) \), denoted as \( C(q(\lambda)) \), is the matrix
\[
\begin{pmatrix}
0 & \cdots & 0 & -a_0 \\
1 & 0 & \cdots & -a_1 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 1 & -a_{n-1}
\end{pmatrix}
\]

**Proposition 7.6.2** Let \( q(\lambda) \) be a polynomial and let \( C(q(\lambda)) \) be its companion matrix. Then \( q(C(q(\lambda))) = 0 \).

**Proof:** Write \( C \) instead of \( C(q(\lambda)) \) for short. Note that
\[ Ce_1 = e_2, Ce_2 = e_3, \ldots, Ce_{n-1} = e_n \]
Thus
\[ e_k = C^{k-1}e_1, \ k = 1, \ldots, n \]
and so it follows
\[ \{e_1, Ce_1, C^2e_1, \ldots, C^{n-1}e_1\} \]
are linearly independent. Hence these form a basis for \( \mathbb{F}^n \). Now note that \( Ce_n \) is given by
\[ Ce_n = -a_0e_1 - a_1e_2 - \cdots - a_{n-1}e_n \]
and from (7.10) this implies
\[ C^n e_1 = -a_0e_1 - a_1Ce_1 - \cdots - a_{n-1}C^{n-1}e_1 \]
and so \( q(C)e_1 = 0 \). Now since (7.11) is a basis, every vector of \( \mathbb{F}^n \) is of the form \( k(C)e_1 \) for some polynomial \( k(\lambda) \). Therefore, if \( v \in \mathbb{F}^n \),
\[ q(C)v = q(C)k(C)e_1 = k(C)q(C)e_1 = 0 \]
which shows \( q(C) = 0 \).

In the case where two \( n \times n \) matrices \( M, N \) are similar, recall this is equivalent to the two being matrices of the same linear transformation taken with respect to two different bases. Hence each are similar to the same rational canonical form.

**Example 7.6.3** Here is a matrix.
\[
A = \begin{pmatrix}
5 & -2 & 1 \\
2 & 10 & -2 \\
9 & 0 & 9
\end{pmatrix}
\]
Find a similarity transformation which will produce the rational canonical form for \( A \).

The minimum polynomial is \( \lambda^3 - 24\lambda^2 + 180\lambda - 432 \). This factors as
\[ (\lambda - 6)^2(\lambda - 12) \]
Thus $\mathbb{Q}^3$ is the direct sum of $\ker \left( (A - 6I)^2 \right)$ and $\ker (A - 12I)$. Consider the first of these. You see easily that this is

$$y \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, y, z \in \mathbb{Q}.$$ 

What about the length of $A$ cyclic sets? It turns out it doesn’t matter much. You can start with either of these and get a cycle of length 2. Let’s pick the second one. This leads to the cycle

$$\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -4 \\ -4 \\ 1 \end{pmatrix} = A \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -12 \\ -48 \\ -36 \end{pmatrix} = A^2 \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix},$$

where the last of the three is a linear combination of the first two. Take the first two as the first two columns of $S$. To get the third, you need a cycle of length 1 corresponding to $\ker (A - 12I)$. This yields the eigenvector $\begin{pmatrix} 1 & -2 & 3 \end{pmatrix}^T$. Thus

$$S = \begin{pmatrix} -1 & -4 & 1 \\ 0 & -4 & -2 \\ 1 & 0 & 3 \end{pmatrix}.$$ 

Now using Proposition 5.2.10, the Rational canonical form for $A$ should be

$$\begin{pmatrix} -1 & -4 & 1 \\ 0 & -4 & -2 \\ 1 & 0 & 3 \end{pmatrix}^{-1} \begin{pmatrix} 5 & -2 & 1 \\ 2 & 10 & -2 \\ 9 & 0 & 9 \end{pmatrix} \begin{pmatrix} -1 & -4 & 1 \\ 0 & -4 & -2 \\ 1 & 0 & 3 \end{pmatrix} = \begin{pmatrix} 0 & -36 & 0 \\ 1 & 12 & 0 \\ 0 & 0 & 12 \end{pmatrix}.$$ 

Note that $(\lambda - 6)^2 = \lambda^2 - 12\lambda + 36$ and so the top left block is indeed the companion matrix of this polynomial. Of course in this case, we could have obtained the Jordan canonical form.

**Example 7.6.4** Here is a matrix.

$$A = \begin{pmatrix} 12 & -3 & -19 & -14 & 8 \\ -4 & 1 & 1 & 6 & -4 \\ 4 & 5 & 5 & -2 & 4 \\ 0 & -5 & -5 & 2 & 0 \\ -4 & 3 & 11 & 6 & 0 \end{pmatrix}$$

Find a basis such that if $S$ is the matrix which has these vectors as columns $S^{-1}AS$ is in rational canonical form assuming the field of scalars is $\mathbb{Q}$.

The minimum polynomial is

$$\lambda^3 - 12\lambda^2 + 64\lambda - 128$$

This polynomial factors as

$$(\lambda - 4) \left( \lambda^2 - 8\lambda + 32 \right) \equiv \phi_1 (\lambda) \phi_2 (\lambda)$$

where the second factor is irreducible over $\mathbb{Q}$. Consider $\phi_2 (\lambda)$ first. Messy computations yield

$$\ker (\phi_2 (A)) = a \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + b \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} + c \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} + d \begin{pmatrix} -2 \\ 0 \\ 1 \end{pmatrix},$$
Now start with one of these basis vectors and look for an $A$ cycle. Picking the first one, you obtain the cycle

$$
\begin{pmatrix}
-1 \\
1 \\
0 \\
0 \\
0
\end{pmatrix},
\begin{pmatrix}
-15 \\
5 \\
1 \\
-5 \\
7
\end{pmatrix}
$$

because the next vector involving $A^2$ yields a vector which is in the span of the above two. You check this by making the vectors the columns of a matrix and finding the row reduced echelon form. Clearly this cycle does not span $\ker(\phi_2(A))$, so look for another cycle. Begin with a vector which is not in the span of these two. The last one works well. Thus another $A$ cycle is

$$
\begin{pmatrix}
-2 \\
0 \\
0 \\
1
\end{pmatrix},
\begin{pmatrix}
-16 \\
4 \\
-4 \\
8
\end{pmatrix}
$$

It follows a basis for $\ker(\phi_2(A))$ is

$$
\left\{ \begin{pmatrix}
-2 \\
0 \\
0 \\
1
\end{pmatrix},
\begin{pmatrix}
-16 \\
4 \\
-4 \\
8
\end{pmatrix},
\begin{pmatrix}
-1 \\
1 \\
0 \\
0
\end{pmatrix},
\begin{pmatrix}
-15 \\
5 \\
1 \\
-5
\end{pmatrix} \right\}
$$

Finally consider a cycle coming from $\ker(\phi_1(A))$. This amounts to nothing more than finding an eigenvector for $A$ corresponding to the eigenvalue 4. An eigenvector is

$$
\begin{pmatrix}
-1 \\
0 \\
0 \\
1
\end{pmatrix}^T
$$

Now the desired matrix for the similarity transformation is

$$
S =
\begin{pmatrix}
-2 & -16 & -1 & -15 & -1 \\
0 & 4 & 1 & 5 & 0 \\
0 & -4 & 0 & 1 & 0 \\
0 & 0 & 0 & -5 & 0 \\
1 & 8 & 0 & 7 & 1
\end{pmatrix}
$$

Then doing the computations, you get

$$
S^{-1}AS =
\begin{pmatrix}
0 & -32 & 0 & 0 & 0 \\
-1 & 8 & 0 & 0 & 0 \\
0 & 0 & 0 & -32 & 0 \\
0 & 0 & 1 & 8 & 0 \\
0 & 0 & 0 & 0 & 4
\end{pmatrix}
$$

and you see this is in rational canonical form, the two $2 \times 2$ blocks being companion matrices for the polynomial $\lambda^2 - 8\lambda + 32$ and the $1 \times 1$ block being a companion matrix for $\lambda - 4$. Note that you could have written this without finding a similarity transformation to produce it. This follows from the above theory which gave the existence of the rational canonical form.
Obviously there is a lot more which could be considered about rational canonical forms. Just begin with a strange field and start investigating what can be said. One can also derive more systematic methods for finding the rational canonical form. The advantage of this is you don’t need to find the eigenvalues in order to compute the rational canonical form and it can often be computed for this reason, unlike the Jordan form. The uniqueness of this rational canonical form can be used to determine whether two matrices consisting of entries in some field are similar, provided you are able to factor the minimum polynomial into the product of irreducible factors.

### 7.7 Exercises

1. Find the minimum polynomial for

\[ A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ -3 & 2 & 1 \end{pmatrix} \]

assuming the field of scalars is the rational numbers.

2. Show, using the rational root theorem, the minimum polynomial for \( A \) in the above problem is irreducible with respect to \( \mathbb{Q} \). Letting the field of scalars be \( \mathbb{Q} \) find the rational canonical form and a similarity transformation which will produce it.

3. Letting the field of scalars be \( \mathbb{Q} \), find the rational canonical form for the matrix

\[ \begin{pmatrix} 1 & 2 & 1 & -1 \\ 2 & 3 & 0 & 2 \\ 1 & 3 & 2 & 4 \\ 1 & 2 & 1 & 2 \end{pmatrix} \]

4. Let \( A : \mathbb{Q}^3 \to \mathbb{Q}^3 \) be linear. Suppose the minimum polynomial is \( (\lambda - 2) (\lambda^2 + 2\lambda + 7) \). Find the rational canonical form. Can you give generalizations of this rather simple problem to other situations?

5. Find the rational canonical form with respect to the field of scalars equal to \( \mathbb{Q} \) for the matrix

\[ A = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{pmatrix} \]

Observe that this particular matrix is already a companion matrix of \( \lambda^3 - \lambda^2 + \lambda - 1 \). Then find the rational canonical form if the field of scalars equals \( \mathbb{C} \) or \( \mathbb{Q} + i\mathbb{Q} \).

6. Suppose you have two \( n \times n \) matrices \( A, B \) whose entries are in a field \( \mathbb{F} \) and suppose \( \mathbb{G} \) is an extension of \( \mathbb{F} \). For example, you could have \( \mathbb{F} = \mathbb{Q} \) and \( \mathbb{G} = \mathbb{C} \). Suppose \( A \) and \( B \) are similar with respect to the field \( \mathbb{G} \). Can it be concluded that they are similar with respect to the field \( \mathbb{F} \)? **Hint:** First show that the two have the same minimum polynomial over \( \mathbb{F} \). Let this be \( \prod_{i=1}^{q} \phi_i (\lambda)^{p_i} \). Say \( \beta_{v_j} \) is associated with the polynomial \( \phi (\lambda)^{p_j} \). Thus, as described above \( |\beta_{v_j}| \) equals \( p_j d \). Consider the following table which comes from the \( A \) cyclic set

\[ \{ v_j, Av_j, \cdots, A^{d-1}v_j, \cdots, A^{p_j d - 1}v_j \} \]
In the above, \( \alpha_k^j \) signifies the vectors below it in the \( k \)th column. None of these vectors below the top row are equal to 0 because the degree of \( \phi(A)^{p_j-1} \lambda^{d-1} \) is \( dp_j - 1 \), which is less than \( p_j d \) and the smallest degree of a nonzero polynomial sending \( v_j \) to 0 is \( p_j d \). Thus the vector on the bottom of each column is an eigenvector for \( \phi(A) \). Also, each of these vectors is in the span of \( \beta_{v_j} \) and there are \( dp_j \) of them, just as there are \( dp_j \) vectors in \( v_j \). Now show that \( \{ \alpha_0^j, \cdots, \alpha_{d-1}^j \} \) is independent. If you string these together in opposite order, you get a basis which yields a block composed of sub-blocks for the Jordan canonical form for \( \phi(A) \). Now show that \( \phi(A), \phi(B) \) are similar with respect to \( G \). Therefore, they have exactly the same Jordan canonical form with respect to \( G \). The lengths of the cycles are determined according to the size of these Jordan blocks and so you must have the same lengths of cycles for \( A \) as you do for \( B \) which shows they must have the same rational canonical form. Therefore, they are similar with respect to \( F \). Thus all you have to do is to verify that the set is independent and that \( \phi(A), \phi(B) \) are similar with respect to \( G \).

To show why the set is independent, suppose

\[
\sum_{i=0}^{d-1} \sum_{k=0}^{p_j-1} c_{ik} \phi(A)^i A^k v_j = 0
\]

Then multiplying both sides by \( \phi(A)^{p_j-1} \). Explain why

\[
\sum_{i=0}^{d-1} c_{i0} \phi(A)^{p_j-1} A^i v_j = 0
\]

Now if any of the \( c_{i0} \) is nonzero this would imply there exists a polynomial having degree smaller than \( p_j d \) which sends \( v_j \) to 0. This zeros out the \( c_{i0} \) scalars. Next multiply by \( \phi(A)^{p_j-2} \) and give a similar argument. This kind of argument is used by Friedberg Insel and Spence [13] to verify uniqueness of the rational canonical form. Note that this is a very interesting result, being even better than the theorem which gives uniqueness of the rational canonical form with respect to a single field.
Chapter 8

Determinants

The determinant is a number which comes from an $n \times n$ matrix of elements of a field $\mathbb{F}$. It is easiest to give a definition of the determinant which is clearly well defined and then prove the one which involves Laplace expansion which the reader might have seen already. Let $(i_1, \cdots, i_n)$ be an ordered list of numbers from $\{1, \cdots, n\}$. This means the order is important so $(1, 2, 3)$ and $(2, 1, 3)$ are different. Two books which give a good introduction to determinants are Apostol [1] and Rudin [28]. Some recent books which also have a good introduction are Baker [3], and Baker and Kuttler [5]. The approach here is less elegant than in these other books but it amounts to the same thing. I have just tried to avoid the language of permutations in the presentation. The function $\text{sgn}$ presented in what follows is really the sign of a permutation however.

8.1 The Function $\text{sgn}$

The following Lemma will be essential in the definition of the determinant.

Lemma 8.1.1 There exists a function, $\text{sgn}_n$ which maps each ordered list of numbers from $\{1, \cdots, n\}$ to one of the three numbers, 0, 1, or $-1$ which also has the following properties.

\[ \text{sgn}_n (1, \cdots, n) = 1 \]  
\[ \text{sgn}_n (i_1, \cdots, p, \cdots, q, \cdots, i_n) = - \text{sgn}_n (i_1, \cdots, q, \cdots, p, \cdots, i_n) \] (8.2)

In words, the second property states that if two of the numbers are switched, the value of the function is multiplied by $-1$.

Proof: Define $\text{sgn}(x) = 1$ if $x > 0$, $-1$ if $x < 0$ and 0 if $x = 0$. If $n = 1$, there is only one list and it is just the number 1. Thus one can define $\text{sgn}_1 (1) = 1$. For the general case where $n > 1$, simply define

\[ \text{sgn}_n (i_1, \cdots, i_n) = \text{sgn}(\prod_{r<s} (i_s - i_r)) \]  
\[ (-1)^{n-\theta} \text{sgn}_{n-1} (i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n) \] (8.3)

where $n = \theta$ in the ordered list, $(i_1, \cdots, i_n)$.

This delivers either $-1, 1,$ or $0$ by definition. What about the other claims? Suppose you switch $i_p$ with $i_q$ where $p < q$ so two numbers in the ordered list $(i_1, \cdots, i_n)$ are switched. Denote the new ordered list of numbers as $(j_1, \cdots, j_n)$. Thus $j_p = i_q$ and $j_q = i_p$ and if $r \notin \{p,q\}$, $j_r = i_r$. See the following illustration
gives both functions

When you have an ordered list of distinct numbers from 1, \ldots, n:

\[ \sgn_n(i_1, \ldots, i_n) = \text{sign} \left( \prod_{r<s} (j_s - j_r) \right) \]

\[ = \text{sign} \left( \prod_{p<q} (i_p - i_q) \prod_{p<j<q} (i_p - i_j) \prod_{r<s, r \neq p,q} (i_s - i_r) \right) \]

The last product consists of the product of terms which were in \(\prod_{r<s} (i_s - i_r)\) while the two products in the middle both introduce \(q - p - 1\) minus signs. Thus their product is positive. The first factor is of opposite sign to the \(i_q - i_p\) which occurred in \(\sgn_n(i_1, \ldots, i_n)\). Therefore, this switch introduced a minus sign and

\[ \sgn_n(j_1, \ldots, j_n) = -\sgn_n(i_1, \ldots, i_n) \]

Now consider the last claim. In computing \(\sgn_n(i_1, \ldots, i_{\theta-1}, n, i_{\theta+1}, \ldots, i_n)\) there will be the product of \(n-\theta\) negative terms

\[ (i_{\theta+1} - n) \cdots (i_n - n) \]

and the other terms in the product for computing \(\sgn_n(i_1, \ldots, i_{\theta-1}, n, i_{\theta+1}, \ldots, i_n)\) are those which are required to compute \(\sgn_{n-1}(i_1, \ldots, i_{\theta-1}, i_{\theta+1}, \ldots, i_n)\) multiplied by terms of the form \((n - i_j)\) which are nonnegative. It follows that

\[ \sgn_n(i_1, \ldots, i_{\theta-1}, n, i_{\theta+1}, \ldots, i_n) = (-1)^{n-\theta} \sgn_{n-1}(i_1, \ldots, i_{\theta-1}, i_{\theta+1}, \ldots, i_n) \]

It is obvious that if there are repeats in the list the function gives 0.

**Lemma 8.1.2** Every ordered list of distinct numbers from \(\{1, 2, \ldots, n\}\) can be obtained from every other ordered list of distinct numbers by a finite number of switches. Also, \(\sgn_n\) is unique.

**Proof:** This is obvious if \(n = 1\) or 2. Suppose then that it is true for sets of \(n - 1\) elements. Take two ordered lists of numbers, \(P_1, P_2\). Make one switch in both to place \(n\) at the end. Call the result \(P_1^0\) and \(P_2^0\). Then using induction, there are finitely many switches in \(P_1^0\) so that it will coincide with \(P_2^0\). Now switch the \(n\) in what results to where it was in \(P_2\).

To see \(\sgn_n\) is unique, if there exist two functions, \(f\) and \(g\) both satisfying \(\Box\) and \(\Box\) you could start with \(f(1, \ldots, n) = g(1, \ldots, n) = 1\) and applying the same sequence of switches, eventually arrive at \(f(1, \ldots, n) = g(1, \ldots, n)\). If any numbers are repeated, then \(\Box\) gives both functions are equal to zero for that ordered list.

**Definition 8.1.3** When you have an ordered list of distinct numbers from \(\{1, 2, \ldots, n\}\), say

\( (i_1, \ldots, i_n) \),

this ordered list is called a permutation. The symbol for all such permutations is \(S_n\). The number \(\sgn_n(i_1, \ldots, i_n)\) is called the sign of the permutation.

A permutation can also be considered as a function from the set \(\{1, 2, \ldots, n\}\) to \(\{1, 2, \ldots, n\}\) as follows. Let \(f(k) = i_k\). Permutations are of fundamental importance in certain areas of math. For example, it was by considering permutations that Galois was able to give a criterion for solution of polynomial equations by radicals, but this is a different direction than what is being attempted here.

In what follows \(\sgn\) will often be used rather than \(\sgn_n\) because the context supplies the appropriate \(n\).
8.2 The Definition Of The Determinant

Definition 8.2.1 Let \( f \) be a real valued function which has the set of ordered lists of numbers from \( \{1, \ldots, n\} \) as its domain. Define

\[
\sum_{(k_1, \ldots, k_n)} f(k_1 \cdots k_n)
\]

to be the sum of all the \( f(k_1 \cdots k_n) \) for all possible choices of ordered lists \( (k_1, \ldots, k_n) \) of numbers of \( \{1, \ldots, n\} \). For example,

\[
\sum_{(k_1, k_2)} f(k_1, k_2) = f(1, 2) + f(2, 1) + f(1, 1) + f(2, 2).
\]

Definition 8.2.2 Let \( (a_{ij}) = A \) denote an \( n \times n \) matrix. The determinant of \( A \), denoted by \( \det(A) \), is defined by

\[
\det(A) \equiv \sum_{(k_1, \ldots, k_n)} \text{sgn}(k_1, \ldots, k_n) a_{1k_1} \cdots a_{nk_n}
\]

where the sum is taken over all ordered lists of numbers from \( \{1, \ldots, n\} \). Note it suffices to take the sum over only those ordered lists in which there are no repeats because if there are, \( \text{sgn}(k_1, \ldots, k_n) = 0 \) and so that term contributes 0 to the sum.

Let \( A \) be an \( n \times n \) matrix \( A = (a_{ij}) \) and let \( (r_1, \ldots, r_n) \) denote an ordered list of \( n \) numbers from \( \{1, \ldots, n\} \). Let \( A(r_1, \ldots, r_n) \) denote the matrix whose \( k^{th} \) row is the \( r_k \) row of the matrix \( A \). Thus

\[
\det(A(r_1, \ldots, r_n)) = \sum_{(k_1, \ldots, k_n)} \text{sgn}(k_1, \ldots, k_n) a_{r_1k_1} \cdots a_{r_nk_n}
\]

and \( A(1, \ldots, n) = A \).

Proposition 8.2.3 Let \( (r_1, \ldots, r_n) \) be an ordered list of numbers from \( \{1, \ldots, n\} \). Then

\[
\text{sgn}(r_1, \ldots, r_n) \det(A) = \sum_{(k_1, \ldots, k_n)} \text{sgn}(k_1, \ldots, k_n) a_{r_1k_1} \cdots a_{r_nk_n}
\]

and renaming the variables, calling \( k_s, k_r \) and \( k_r, k_s \), this equals

\[
\sum_{(k_1, \ldots, k_n)} \text{sgn}(k_1, \ldots, k_r, \ldots, k_s, \ldots, k_n) a_{1k_1} \cdots a_{rk_r} \cdots a_{sk_s} \cdots a_{nk_n},
\]

and

\[
= \sum_{(k_1, \ldots, k_n)} \text{sgn}(k_1, \ldots, k_r, \ldots, k_s, \ldots, k_n) a_{1k_1} \cdots a_{rk_r} \cdots a_{sk_s} \cdots a_{nk_n}
\]

These got switched

\[
= -\text{det}(A(1, \ldots, s, \ldots, r, \ldots, n)).
\]

Consequently,

\[
\det(A(1, \ldots, s, \ldots, r, \ldots, n)) = -\det(A(1, \ldots, r, \ldots, s, \ldots, n)) = -\det(A)
\]
Now letting $A (1, \cdots , s, \cdots , r, \cdots , n)$ play the role of $A$, and continuing in this way, switching pairs of numbers,

$$\det (A (r_1, \cdots , r_n)) = (-1)^p \det (A)$$

where it took $p$ switches to obtain $(r_1, \cdots , r_n)$ from $(1, \cdots , n)$. By Lemma 8.1.1, this implies

$$\det (A (r_1, \cdots , r_n)) = (-1)^p \det (A) = \sgn (r_1, \cdots , r_n) \det (A)$$

and proves the proposition in the case when there are no repeated numbers in the ordered list, $(r_1, \cdots , r_n)$. However, if there is a repeat, say the $r^\text{th}$ row equals the $s^\text{th}$ row, then the reasoning of 8.2.3 shows that $\det (A (r_1, \cdots , r_n)) = 0$ and also $\sgn (r_1, \cdots , r_n) = 0$ so the formula holds in this case also. ■

**Observation 8.2.4** There are $n!$ ordered lists of distinct numbers from $\{1, \cdots , n\}$.

To see this, consider $n$ slots placed in order. There are $n$ choices for the first slot. For each of these choices, there are $n-1$ choices for the second. Thus there are $n(n-1)$ ways to fill the first two slots. Then for each of these ways there are $n-2$ choices left for the third slot. Continuing this way, there are $n!$ ordered lists of distinct numbers from $\{1, \cdots , n\}$ as stated in the observation.

### 8.3 A Symmetric Definition

With the above, it is possible to give a more symmetric description of the determinant from which it will follow that $\det (A) = \det (A^T)$.

**Corollary 8.3.1** The following formula for $\det (A)$ is valid.

$$\det (A) = \frac{1}{n!} \sum_{(r_1, \cdots , r_n)} \sum_{(k_1, \cdots , k_n)} \sgn (r_1, \cdots , r_n) \sgn (k_1, \cdots , k_n) a_{r_1k_1} \cdots a_{r_nk_n}. \quad (8.9)$$

And also $\det (A^T) = \det (A)$ where $A^T$ is the transpose of $A$. (Recall that for $A^T = (a_{ij}^T)$, $a_{ij}^T = a_{ji}$.)

**Proof:** From Proposition 8.2.3, if the $r_i$ are distinct,

$$\det (A) = \sum_{(k_1, \cdots , k_n)} \sgn (r_1, \cdots , r_n) \sgn (k_1, \cdots , k_n) a_{r_1k_1} \cdots a_{r_nk_n}.$$ 

Summing over all ordered lists, $(r_1, \cdots , r_n)$ where the $r_i$ are distinct, (If the $r_i$ are not distinct, $\sgn (r_1, \cdots , r_n) = 0$ and so there is no contribution to the sum.)

$$n! \det (A) = \sum_{(r_1, \cdots , r_n)} \sum_{(k_1, \cdots , k_n)} \sgn (r_1, \cdots , r_n) \sgn (k_1, \cdots , k_n) a_{r_1k_1} \cdots a_{r_nk_n}. $$

This proves the corollary since the formula gives the same number for $A$ as it does for $A^T$. ■

**Corollary 8.3.2** If two rows or two columns in an $n \times n$ matrix $A$, are switched, the determinant of the resulting matrix equals $(-1)$ times the determinant of the original matrix. If $A$ is an $n \times n$ matrix in which two rows are equal or two columns are equal then $\det (A) = 0$. Suppose the $i^\text{th}$ row of $A$ equals $(xa_1 + yb_1, \cdots , xa_n + yb_n)$. Then

$$\det (A) = x \det (A_1) + y \det (A_2)$$

where the $i^\text{th}$ row of $A_1$ is $(a_1, \cdots , a_n)$ and the $i^\text{th}$ row of $A_2$ is $(b_1, \cdots , b_n)$, all other rows of $A_1$ and $A_2$ coinciding with those of $A$. In other words, $\det$ is a linear function of each row $A$. The same is true with the word “row” replaced with the word “column”.


8.4. BASIC PROPERTIES OF THE DETERMINANT

**Proof:** By Proposition 8.2.3 when two rows are switched, the determinant of the resulting matrix is $(-1)$ times the determinant of the original matrix. By Corollary 8.3.1 the same holds for columns because the columns of the matrix equal the rows of the transposed matrix. Thus if $A_1$ is the matrix obtained from $A$ by switching two columns,

$$\det (A) = \det (A^T) = - \det (A_1^T) = - \det (A_1).$$

If $A$ has two equal columns or two equal rows, then switching them results in the same matrix. Therefore, $\det (A) = - \det (A)$ and so $\det (A) = 0$.

It remains to verify the last assertion.

$$\det (A) = \sum_{(k_1, \ldots, k_n)} \text{sgn} (k_1, \ldots, k_n) a_{1k_1} \cdots (xa_{rk_i} + yb_{rk_i}) \cdots a_{nk_n}$$

$$= x \sum_{(k_1, \ldots, k_n)} \text{sgn} (k_1, \ldots, k_n) a_{1k_1} \cdots a_{rk_i} \cdots a_{nk_n}$$

$$+ y \sum_{(k_1, \ldots, k_n)} \text{sgn} (k_1, \ldots, k_n) a_{1k_1} \cdots b_{rk_i} \cdots a_{nk_n} = x \det (A_1) + y \det (A_2).$$

The same is true of columns because $\det (A^T) = \det (A)$ and the rows of $A^T$ are the columns of $A$.

---

8.4 Basic Properties Of The Determinant

**Definition 8.4.1** A vector, $w$, is a linear combination of the vectors $\{v_1, \ldots, v_r\}$ if there exist scalars $c_1, \ldots, c_r$ such that $w = \sum_{k=1}^{r} c_k v_k$. This is the same as saying $w \in \text{span} (v_1, \ldots, v_r)$.

The following corollary is also of great use.

**Corollary 8.4.2** Suppose $A$ is an $n \times n$ matrix and some column (row) is a linear combination of $r$ other columns (rows). Then $\det (A) = 0$.

**Proof:** Let $A = \left( \begin{array}{ccc} a_1 & \cdots & a_n \end{array} \right)$ be the columns of $A$ and suppose the condition that one column is a linear combination of $r$ of the others is satisfied. Say $a_i = \sum_{j \neq i} c_j a_j$. Then by Corollary 8.3.2

$$\det (A) = \det \left( \begin{array}{ccc} a_1 & \cdots & \sum_{j \neq i} c_j a_j & \cdots & a_n \end{array} \right) = \sum_{j \neq i} c_j \det \left( \begin{array}{ccc} a_1 & \cdots & a_j & \cdots & a_n \end{array} \right) = 0$$

because each of these determinants in the sum has two equal rows.

Recall the following definition of matrix multiplication.

**Definition 8.4.3** If $A$ and $B$ are $n \times n$ matrices, $A = (a_{ij})$ and $B = (b_{ij})$, $AB = (c_{ij})$ where $c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$.

One of the most important rules about determinants is that the determinant of a product equals the product of the determinants.

**Theorem 8.4.4** Let $A$ and $B$ be $n \times n$ matrices. Then

$$\det (AB) = \det (A) \det (B).$$
**Proof:** Let \( c_{ij} \) be the \( ij^{th} \) entry of \( AB \). Then by Proposition \( \ref{prop:det_product} \)

\[
\det(AB) = \sum_{(k_1, \cdots, k_n)} \text{sgn}(k_1, \cdots, k_n) c_{1k_1} \cdots c_{nk_n}
\]

\[
= \sum_{(k_1, \cdots, k_n)} \text{sgn}(k_1, \cdots, k_n) \left( \sum_{i_1} a_{1i_1} b_{i_1k_1} \right) \cdots \left( \sum_{i_n} a_{ni_n} b_{i_nk_n} \right)
\]

\[
= \sum_{(r_1, \cdots, r_n)} \sum_{(k_1, \cdots, k_n)} \text{sgn}(k_1, \cdots, k_n) b_{r_1k_1} \cdots b_{r_nk_n} (a_{1r_1} \cdots a_{nr_n})
\]

\[
= \sum_{(r_1, \cdots, r_n)} \text{sgn}(r_1 \cdots r_n) a_{1r_1} \cdots a_{nr_n} \det(B) = \det(A) \det(B)
\]

Note that this shows that if two matrices are similar, then they have the same determinant and also the same characteristic polynomial, \( \det(\lambda I - A) \).

The Binet Cauchy formula is a generalization of the theorem which says the determinant of a product is the product of the determinants. The situation is illustrated in the following picture where \( A, B \) are matrices.

\[
\begin{array}{c}
B \\
A
\end{array}
\]

**Theorem 8.4.5** Let \( A \) be an \( n \times m \) matrix with \( n \geq m \) and let \( B \) be a \( m \times n \) matrix. Also let \( A_i \), \( i = 1, \cdots, C(n, m) \)

be the \( m \times m \) submatrices of \( A \) which are obtained by deleting \( n - m \) rows and let \( B_i \) be the \( m \times m \) submatrices of \( B \) which are obtained by deleting corresponding \( n - m \) columns. Then

\[
\det(AB) = \sum_{k=1}^{C(n, m)} \det(B_k) \det(A_k)
\]

**Proof:** This follows from a computation. By Corollary \ref{cor:det_sub} on Page 102, \( \det(AB) = \)

\[
\frac{1}{m!} \sum_{(i_1, \cdots, i_m)} \sum_{(j_1, \cdots, j_m)} \text{sgn}(i_1 \cdots i_m) \text{sgn}(j_1 \cdots j_m) \det(AB)_{i_1j_1} \det(AB)_{i_2j_2} \cdots \det(AB)_{i_mj_m}
\]

\[
= \frac{1}{m!} \sum_{(i_1, \cdots, i_m)} \sum_{(j_1, \cdots, j_m)} \text{sgn}(i_1 \cdots i_m) \text{sgn}(j_1 \cdots j_m) \cdot
\]

\[
\sum_{r_1=1}^{n} B_{i_1r_1} A_{r_1j_1} \sum_{r_2=1}^{n} B_{i_2r_2} A_{r_2j_2} \cdots \sum_{r_m=1}^{n} B_{i_mr_m} A_{r_mj_m}
\]

Now denote by \( I_k \) one of the subsets of \( \{1, \cdots, n\} \) which has \( m \) elements. Thus there are \( C(n, m) \) of these.

\[
= \sum_{k=1}^{C(n, m)} \sum_{(r_1, \cdots, r_m) = I_k} \frac{1}{m!} \sum_{(i_1, \cdots, i_m)} \sum_{(j_1, \cdots, j_m)} \text{sgn}(i_1 \cdots i_m) \text{sgn}(j_1 \cdots j_m) \cdot
\]

\[
B_{i_1r_1} A_{r_1j_1} B_{i_2r_2} A_{r_2j_2} \cdots B_{i_mr_m} A_{r_mj_m}
\]
Suppose a matrix is of the form
to write

$$\text{det } \begin{bmatrix} 1 \cdots i_m \end{bmatrix} B_{i_1 r_1} B_{i_2 r_2} \cdots B_{i_m r_m}.$$ 

Then

$$\text{det } \begin{bmatrix} j_1 \cdots j_m \end{bmatrix} A_{r_1 j_1} A_{r_2 j_2} \cdots A_{r_m j_m}$$

Deﬁnition 8.5.2

expansion along a row or a column. This will follow from the above deﬁnition of a determinant.

To get the assertion in the ﬁrst case, use Corollary 8.5.1

Now suppose the second case. Then if $k_i \neq n$, the term involving $m_{nk_i}$ in the above expression equals zero. Therefore, the only terms which survive are those for which $\theta = n$ or in other words, those for which $k_n = n$. Therefore, the above expression reduces to

$$a \sum_{(k_1, \cdots, k_{n-1})} \text{sgn}_{n-1} (k_1, \cdots, k_{n-1}) \ m_{1k_1} \cdots m_{nk_n} = a \det (A).$$

To get the assertion in the first case, use Corollary 8.5.1 to write

$$\text{det } (M) = \text{det } (M^T) = \det \begin{bmatrix} A^T & 0 \\ * & a \end{bmatrix} = a \det (A^T) = a \det (A).$$

In terms of the theory of determinants, arguably the most important idea is that of Laplace expansion along a row or a column. This will follow from the above deﬁnition of a determinant.

Deﬁnition 8.5.2 Let $A = (a_{ij})$ be an $n \times n$ matrix. Then a new matrix called the cofactor matrix $\text{cof } (A)$ is deﬁned by $\text{cof } (A) = (c_{ij})$ where to obtain $c_{ij}$ delete the $i^{th}$ row and the $j^{th}$ column of $A$, take the determinant of the $(n-1) \times (n-1)$ matrix which results, (This is called the $ij^{th}$ minor of $A$. ) and then multiply this number by $(-1)^{i+j}$. To make the formulas easier to remember, $\text{cof } (A)_{ij}$ will denote the $ij^{th}$ entry of the cofactor matrix.

8.5 Expansion Using Cofactors

Lemma 8.5.1 Suppose a matrix is of the form

$$M = \begin{bmatrix} A & * \\ 0 & a \end{bmatrix} \text{ or } \begin{bmatrix} A & 0 \\ * & a \end{bmatrix} \quad (8.10)$$

where $a$ is a number and $A$ is an $(n-1) \times (n-1)$ matrix and $*$ denotes either a column or a row having length $n-1$ and the $0$ denotes either a column or a row of length $n-1$ consisting entirely of zeros. Then $\text{det } (M) = a \det (A)$.

Proof: Denote $M$ by $(m_{ij})$. Thus in the ﬁrst case, $m_{nn} = a$ and $m_{ni} = 0$ if $i \neq n$ while in the second case, $m_{nm} = a$ and $m_{in} = 0$ if $i \neq n$. From the deﬁnition of the determinant,

$$\det (M) = \sum_{(k_1, \cdots, k_n)} \text{sgn} (k_1, \cdots, k_n) \ m_{1k_1} \cdots m_{nk_n}$$

Letting $\theta$ denote the position of $n$ in the ordered list, $(k_1, \cdots, k_n)$ then using the earlier conventions used to prove Lemma 8.3.1, $\det (M)$ equals

$$\sum_{(k_1, \cdots, k_n)} (-1)^{n-\theta} \text{sgn}_{n-1} \begin{bmatrix} k_1, \cdots, k_{\theta-1}, k_{\theta+1}, \cdots, k_n \end{bmatrix} \ m_{1k_1} \cdots m_{nk_n}$$

Now suppose the second case. Then if $k_n \neq n$, the term involving $m_{nk_n}$ in the above expression equals zero. Therefore, the only terms which survive are those for which $\theta = n$ or in other words, those for which $k_n = n$. Therefore, the above expression reduces to

$$a \sum_{(k_1, \cdots, k_{n-1})} \text{sgn}_{n-1} (k_1, \cdots, k_{n-1}) \ m_{1k_1} \cdots m_{(n-1)k_{n-1}} = a \det (A).$$

To get the assertion in the ﬁrst case, use Corollary 8.5.1 to write

$$\det (M) = \det (M^T) = \det \begin{bmatrix} A^T & 0 \\ * & a \end{bmatrix} = a \det (A^T) = a \det (A).$$


The following is the main result.

**Theorem 8.5.3** Let $A$ be an $n \times n$ matrix where $n \geq 2$. Then

$$\det(A) = \sum_{j=1}^{n} a_{ij} \cof(A)_{ij} = \sum_{i=1}^{n} a_{ij} \cof(A)_{ij}. \quad (8.11)$$

The first formula consists of expanding the determinant along the $i^{th}$ row and the second expands the determinant along the $j^{th}$ column.

**Proof:** Let $(a_{i1}, \ldots, a_{in})$ be the $i^{th}$ row of $A$. Let $B_j$ be the matrix obtained from $A$ by leaving every row the same except the $i^{th}$ row which in $B_j$ equals $(0, \ldots, 0, a_{ij}, 0, \ldots, 0)$. Then by Corollary 8.3.2,

$$\det(A) = \sum_{j=1}^{n} \det(B_j)$$

For example if

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ h & i & j \end{pmatrix}$$

and $i = 2$, then

$$B_1 = \begin{pmatrix} a & b & c \\ d & 0 & 0 \\ h & i & j \end{pmatrix}, B_2 = \begin{pmatrix} a & b & c \\ 0 & e & 0 \\ h & i & j \end{pmatrix}, B_3 = \begin{pmatrix} a & b & c \\ 0 & 0 & f \\ h & i & j \end{pmatrix}$$

Denote by $A^{ij}$ the $(n-1) \times (n-1)$ matrix obtained by deleting the $i^{th}$ row and the $j^{th}$ column of $A$. Thus $\cof(A)_{ij} = (-1)^{i+j} \det(A^{ij})$. At this point, recall that from Proposition 8.2.3, when two rows or two columns in a matrix $M$, are switched, this results in multiplying the determinant of the old matrix by $-1$ to get the determinant of the new matrix. Therefore, by Lemma 8.5.1,

$$\det(B_j) = (-1)^{n-j} (-1)^{n-i} \det\left( \begin{pmatrix} A^{ij} & * \\ 0 & a_{ij} \end{pmatrix} \right)$$

$$= (-1)^{i+j} \det\left( \begin{pmatrix} A^{ij} & * \\ 0 & a_{ij} \end{pmatrix} \right) = a_{ij} \cof(A)_{ij}.$$ 

Therefore,

$$\det(A) = \sum_{j=1}^{n} a_{ij} \cof(A)_{ij}$$

which is the formula for expanding $\det(A)$ along the $i^{th}$ row. Also,

$$\det(A) = \det(A^T) = \sum_{j=1}^{n} a_{ij}^T \cof(A^T)_{ij} = \sum_{j=1}^{n} a_{ji} \cof(A)_{ji}$$

which is the formula for expanding $\det(A)$ along the $i^{th}$ column. \square

### 8.6 A Formula For The Inverse

Note that this gives an easy way to write a formula for the inverse of an $n \times n$ matrix.
8.6. A FORMULA FOR THE INVERSE

Theorem 8.6.1 $A^{-1}$ exists if and only if $\det(A) \neq 0$. If $\det(A) \neq 0$, then $A^{-1} = (a_{ij}^{-1})$ where

$$a_{ij}^{-1} = \det(A)^{-1} \text{cof}(A)_{ji}$$

for cof$(A)_{ij}$ the $ij$th cofactor of $A$.

**Proof:** By Theorem 8.5.3 and letting $(a_{ir}) = A$, if $\det(A) \neq 0$,

$$\sum_{i=1}^{n} a_{ir} \text{cof}(A)_{ir} \det(A)^{-1} = \det(A) \det(A)^{-1} = 1.$$ 

Now in the matrix $A$, replace the $k$th column with the $r$th column and then expand along the $k$th column. This yields for $k \neq r$,

$$\sum_{i=1}^{n} a_{ir} \text{cof}(A)_{ik} \det(A)^{-1} = 0$$

because there are two equal columns by Corollary 8.3.2. Summarizing,

$$\sum_{i=1}^{n} a_{ir} \text{cof}(A)_{ik} \det(A)^{-1} = \delta_{rk}.$$ 

Using the other formula in Theorem 8.5.3 and similar reasoning,

$$\sum_{j=1}^{n} a_{rj} \text{cof}(A)_{kj} \det(A)^{-1} = \delta_{rk}.$$ 

This proves that if $\det(A) \neq 0$, then $A^{-1}$ exists with $A^{-1} = (a_{ij}^{-1})$, where

$$a_{ij}^{-1} = \text{cof}(A)_{ji} \det(A)^{-1}.$$ 

Now suppose $A^{-1}$ exists. Then by Theorem 8.4.4,

$$1 = \det(I) = \det(AA^{-1}) = \det(A) \det(A^{-1})$$

so $\det(A) \neq 0$. ■

The next corollary points out that if an $n \times n$ matrix $A$ has a right or a left inverse, then it has an inverse.

**Corollary 8.6.2** Let $A$ be an $n \times n$ matrix and suppose there exists an $n \times n$ matrix $B$ such that $BA = I$. Then $A^{-1}$ exists and $A^{-1} = B$. Also, if there exists $C$ an $n \times n$ matrix such that $AC = I$, then $A^{-1}$ exists and $A^{-1} = C$.

**Proof:** Since $BA = I$, Theorem 8.5.3 implies $\det B \det A = 1$ and so $\det A \neq 0$. Therefore from Theorem 8.6.1, $A^{-1}$ exists. Therefore,

$$A^{-1} = (BA) A^{-1} = B \left( AA^{-1} \right) = BI = B.$$ 

The case where $CA = I$ is handled similarly. ■

The conclusion of this corollary is that left inverses, right inverses and inverses are all the same in the context of $n \times n$ matrices.

Theorem 8.6.1 says that to find the inverse, take the transpose of the cofactor matrix and divide by the determinant. The transpose of the cofactor matrix is called the adjugate or sometimes the classical adjoint of the matrix $A$. It is an abomination to call it the adjoint although you do sometimes see it referred to in this way. In words, $A^{-1}$ is equal to one over the determinant of $A$ times the adjugate matrix of $A$. 


In case you are solving a system of equations, \( Ax = y \) for \( x \), it follows that if \( A^{-1} \) exists,

\[
x = (A^{-1}A)x = A^{-1}(Ax) = A^{-1}y
\]

thus solving the system. Now in the case that \( A^{-1} \) exists, there is a formula for \( A^{-1} \) given above. Using this formula,

\[
x_i = \sum_{j=1}^{n} a_{ij}^{-1} y_j = \sum_{j=1}^{n} \frac{1}{\det(A)} \text{cof}(A)_{ji} y_j.
\]

By the formula for the expansion of a determinant along a column,

\[
x_i = \frac{1}{\det(A)} \begin{vmatrix}
* & \cdots & y_1 & \cdots & *\\
0 & * & \ddots & \vdots \\
\vdots & \ddots & \ddots & * \\
0 & \cdots & 0 & *
\end{vmatrix},
\]

where here the \( i \)th column of \( A \) is replaced with the column vector, \((y_1, \cdots, y_n)^T\), and the determinant of this modified matrix is taken and divided by \( \det(A) \). This formula is known as Cramer’s rule.

**Definition 8.6.3** A matrix \( M \), is upper triangular if \( M_{ij} = 0 \) whenever \( i > j \). Thus such a matrix equals zero below the main diagonal, the entries of the form \( M_{ii} \) as shown.

\[
\begin{pmatrix}
* & * & \cdots & * \\
0 & * & \ddots & \\
\vdots & \ddots & \ddots & * \\
0 & \cdots & 0 & *
\end{pmatrix}
\]

A lower triangular matrix is defined similarly as a matrix for which all entries above the main diagonal are equal to zero.

With this definition, here is a simple corollary of Theorem

**Corollary 8.6.4** Let \( M \) be an upper (lower) triangular matrix. Then \( \det(M) \) is obtained by taking the product of the entries on the main diagonal.

### 8.7 Rank Of A Matrix

**Definition 8.7.1** A submatrix of a matrix \( A \) is the rectangular array of numbers obtained by deleting some rows and columns of \( A \). Let \( A \) be an \( m \times n \) matrix. The determinant rank of the matrix equals \( r \) where \( r \) is the largest number such that some \( r \times r \) submatrix of \( A \) has a non zero determinant. The row rank is defined to be the dimension of the span of the rows. The column rank is defined to be the dimension of the span of the columns.

**Theorem 8.7.2** If \( A \), an \( m \times n \) matrix has determinant rank \( r \), then there exist \( r \) rows of the matrix such that every other row is a linear combination of these \( r \) rows.

**Proof:** Suppose the determinant rank of \( A = (a_{ij}) \) equals \( r \). Thus some \( r \times r \) submatrix has non zero determinant and there is no larger square submatrix which has non zero determinant. Suppose such a submatrix is determined by the \( r \) columns whose indices are

\[
j_1 < \cdots < j_r
\]

and the \( r \) rows whose indices are

\[
i_1 < \cdots < i_r
\]
I want to show that every row is a linear combination of these rows. Consider the \( l \)th row and let \( p \) be an index between 1 and \( n \). Form the following \((r + 1) \times (r + 1)\) matrix

\[
\begin{pmatrix}
    a_{i_1,j_1} & \cdots & a_{i_1,j_r} & a_{i_1,p} \\
    \vdots & & \vdots & \vdots \\
    a_{i_r,j_1} & \cdots & a_{i_r,j_r} & a_{i_r,p} \\
    a_{l_1,j_1} & \cdots & a_{l_r,j_r} & a_{l_1,p}
\end{pmatrix}
\]

Of course you can assume \( l \notin \{i_1, \cdots, i_r\} \) because there is nothing to prove if the \( l \)th row is one of the chosen ones. The above matrix has determinant 0. This is because if \( p \notin \{j_1, \cdots, j_r\} \) then the above would be a submatrix of \( A \) which is too large to have non zero determinant. On the other hand, if \( p \in \{j_1, \cdots, j_r\} \) then the above matrix has two columns which are equal so its determinant is still 0.

Expand the determinant of the above matrix along the last column. Let \( C_k \) denote the cofactor associated with the entry \( a_{i_k,p} \). This is not dependent on the choice of \( p \). Remember, you delete the column and the row the entry is in and take the determinant of what is left and multiply by \(-1\) raised to an appropriate power. Let \( C \) denote the cofactor associated with \( a_{l_1,p} \). This is given to be nonzero, it being the determinant of the matrix \( r \times r \) matrix in the upper left corner. Thus

\[
0 = a_{l_1,p}C + \sum_{k=1}^{r} C_k a_{i_k,p}
\]

which implies

\[
a_{l_1,p} = \sum_{k=1}^{r} \frac{-C_k}{C} a_{i_k,p} = \sum_{k=1}^{r} m_k a_{i_k,p}
\]

Since this is true for every \( p \) and since \( m_k \) does not depend on \( p \), this has shown the \( l \)th row is a linear combination of the \( i_1, i_2, \cdots, i_r \) rows.

**Corollary 8.7.3** The determinant rank equals the row rank.

**Proof:** From Theorem 8.7.2, every row is in the span of \( r \) rows where \( r \) is the determinant rank. Therefore, the row rank (dimension of the span of the rows) is no larger than the determinant rank. Could the row rank be smaller than the determinant rank? If so, it follows from Theorem 8.7.2 that there exist \( p \) rows for \( p < r \equiv \) determinant rank, such that the span of these \( p \) rows equals the row space. But then you could consider the \( r \times r \) sub matrix which determines the determinant rank and it would follow that each of these rows would be in the span of the restrictions of the \( p \) rows just mentioned. By Theorem 3.1.5, the exchange theorem, the rows of this sub matrix would not be linearly independent and so some row is a linear combination of the others. By Corollary 8.4.2 the determinant would be 0, a contradiction.

**Corollary 8.7.4** If \( A \) has determinant rank \( r \), then there exist \( r \) columns of the matrix such that every other column is a linear combination of these \( r \) columns. Also the column rank equals the determinant rank.

**Proof:** This follows from the above by considering \( A^T \). The rows of \( A^T \) are the columns of \( A \) and the determinant rank of \( A^T \) and \( A \) are the same. Therefore, from Corollary 8.7.3, column rank of \( A = \) row rank of \( A^T = \) determinant rank of \( A^T = \) determinant rank of \( A \).

The following theorem is of fundamental importance and ties together many of the ideas presented above.

**Theorem 8.7.5** Let \( A \) be an \( n \times n \) matrix. Then the following are equivalent.

1. \( \det (A) = 0 \).
2. $A, A^T$ are not one to one.

3. $A$ is not onto.

**Proof:** Suppose $\det(A) = 0$. Then the determinant rank of $A = r < n$. Therefore, there exist $r$ columns such that every other column is a linear combination of these columns by Theorem 8.7.2. In particular, it follows that for some $m$, the $m^{th}$ column is a linear combination of all the others. Thus letting $A = \begin{pmatrix} a_1 & \cdots & a_m & \cdots & a_n \end{pmatrix}$ where the columns are denoted by $a_i$, there exists scalars $\alpha_i$ such that

$$a_m = \sum_{k \neq m} \alpha_k a_k.$$

Now consider the column vector, $x \equiv \begin{pmatrix} \alpha_1 & \cdots & -1 & \cdots & \alpha_n \end{pmatrix}^T$. Then

$$Ax = -a_m + \sum_{k \neq m} \alpha_k a_k = 0.$$

Since also $A0 = 0$, it follows $A$ is not one to one. Similarly, $A^T$ is not one to one by the same argument applied to $A^T$. This verifies that 1.) implies 2.).

Now suppose 2.). Then since $A^T$ is not one to one, it follows there exists $x \neq 0$ such that

$$A^T x = 0.$$

Taking the transpose of both sides yields

$$x^T A = 0^T$$

where the $0^T$ is a $1 \times n$ matrix or row vector. Now if $Ay = x$, then

$$|x|^2 = x^T (Ay) = (x^T A) y = 0 y = 0$$

contrary to $x \neq 0$. Consequently there can be no $y$ such that $Ay = x$ and so $A$ is not onto. This shows that 2.) implies 3.).

Finally, suppose 3.). If 1.) does not hold, then $\det(A) \neq 0$ but then from Theorem 8.6.1 $A^{-1}$ exists and so for every $y \in \mathbb{F}^n$ there exists a unique $x \in \mathbb{F}^n$ such that $Ax = y$. In fact $x = A^{-1} y$. Thus $A$ would be onto contrary to 3.). This shows 3.) implies 1.).

**Corollary 8.7.6** Let $A$ be an $n \times n$ matrix. Then the following are equivalent.

1. $\det(A) \neq 0$.
2. $A$ and $A^T$ are one to one.
3. $A$ is onto.

**Proof:** This follows immediately from the above theorem.

### 8.8 Summary Of Determinants

In all the following $A, B$ are $n \times n$ matrices

1. $\det(A)$ is a number.
2. $\det(A)$ is linear in each row and in each column.
3. If you switch two rows or two columns, the determinant of the resulting matrix is \(-1\) times the determinant of the unswitched matrix. (This and the previous one say
\[
(a_1 \cdots a_n) \rightarrow \det (a_1 \cdots a_n)
\]
is an alternating multilinear function or alternating tensor.

4. \(\det (e_1, \cdots, e_n) = 1\).

5. \(\det (AB) = \det (A) \det (B)\)

6. \(\det (A)\) can be expanded along any row or any column and the same result is obtained.

7. \(\det (A) = \det (A^T)\)

8. \(A^{-1}\) exists if and only if \(\det (A) \neq 0\) and in this case
\[
(A^{-1})_{ij} = \frac{1}{\det (A)} \text{cof} (A)_{ji}
\]

9. Determinant rank, row rank and column rank are all the same number for any \(m \times n\) matrix.

8.9 The Cayley Hamilton Theorem

Here is a simple proof of the Cayley Hamilton theorem in the special case that the field of scalars is \(\mathbb{R}, \mathbb{Q},\) or \(\mathbb{C}\). This proof does not work for arbitrary fields. A proof of this theorem valid for every field will be outlined in exercises. See Problem 21 on Page 182. The cases considered here comprise most major applications of the Cayley Hamilton theorem.

Definition 8.9.1 Let \(A\) be an \(n \times n\) matrix. The characteristic polynomial is defined as
\[
q_A (t) \equiv \det (tI - A)
\]
and the solutions to \(q_A (t) = 0\) are called eigenvalues. For \(A\) a matrix and \(p(t) = t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0\), denote by \(p(A)\) the matrix defined by
\[
p(A) \equiv A^n + a_{n-1}A^{n-1} + \cdots + a_1 A + a_0 I.
\]
The explanation for the last term is that \(A^0\) is interpreted as \(I\), the identity matrix. This is always the characteristic polynomial, but in this section, the field will be one of those mentioned above.

The Cayley Hamilton theorem states that every matrix satisfies its characteristic equation, that equation defined by \(q_A (t) = 0\). It is one of the most important theorems in linear algebra.\(^{1}\) The proof in this section is not the most general proof, but works well when the field of scalars is \(\mathbb{R}\) or \(\mathbb{C}\). The following lemma will help with its proof.

Lemma 8.9.2 Suppose for all \(|\lambda|\) large enough,
\[
A_0 + A_1 \lambda + \cdots + A_m \lambda^m = 0,
\]
where the \(A_i\) are \(n \times n\) matrices. Then each \(A_i = 0\).

Proof: Suppose some \(A_i \neq 0\). Let \(p\) be the largest index of those which are non zero. Then multiply by \(\lambda^{-p}\).
\[
A_0 \lambda^{-p} + A_1 \lambda^{-p+1} + \cdots + A_{p-1} \lambda^{-1} + A_p = 0
\]
Now let \(\lambda \rightarrow \infty\). Thus \(A_p = 0\) after all. Hence each \(A_i = 0\). ■

With the lemma, here is a simple corollary.

\(^{1}\)A special case was first proved by Hamilton in 1853. The general case was announced by Cayley some time later and a proof was given by Frobenius in 1878.
Corollary 8.9.3 Let $A_i$ and $B_i$ be $n \times n$ matrices and suppose
\[ A_0 + A_1 \lambda + \cdots + A_m \lambda^n = B_0 + B_1 \lambda + \cdots + B_m \lambda^n \]
for all $|\lambda|$ large enough. Then $A_i = B_i$ for all $i$. If $A_i = B_i$ for each $A_i, B_i$ then one can substitute
an $n \times n$ matrix $M$ for $\lambda$ and the identity will continue to hold.

Proof: Subtract and use the result of the lemma. The last claim is obvious by matching terms.

With this preparation, here is a relatively easy proof of the Cayley Hamilton theorem.

Theorem 8.9.4 Let $A$ be an $n \times n$ matrix and let $q(\lambda) \equiv \det (\lambda I - A)$ be the characteristic polynomial. Then $q(A) = 0$.

Proof: Let $C(\lambda)$ equal the transpose of the cofactor matrix of $(\lambda I - A)$ for $|\lambda|$ large. (If $|\lambda|$ is
large enough, then $\lambda$ cannot be in the finite list of eigenvalues of $A$ and so for such $\lambda$, $(\lambda I - A)^{-1}$
exists.) Therefore, by Theorem 8.9.4
\[ C(\lambda) = q(\lambda) (\lambda I - A)^{-1}. \]

Say
\[ q(\lambda) = a_0 + a_1 \lambda + \cdots + \lambda^n \]
Note that each entry in $C(\lambda)$ is a polynomial in $\lambda$ having degree no more than $n - 1$. For example,
you might have something like
\[ C(\lambda) = \begin{pmatrix} \lambda^2 - 6\lambda + 9 & 3 - \lambda & 0 \\ 2\lambda - 6 & \lambda^2 - 3\lambda & 0 \\ \lambda - 1 & \lambda - 1 & \lambda^2 - 3\lambda + 2 \end{pmatrix} \]
\[ = \begin{pmatrix} 9 & 3 & 0 \\ -6 & 0 & 0 \\ -1 & -1 & 2 \end{pmatrix} + \lambda \begin{pmatrix} -6 & -1 & 0 \\ 2 & -3 & 0 \\ 1 & 1 & -3 \end{pmatrix} + \lambda^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \]
Therefore, collecting the terms in the general case,
\[ C(\lambda) = C_0 + C_1 \lambda + \cdots + C_{n-1} \lambda^{n-1} \]
for $C_j$ some $n \times n$ matrix. Then
\[ C(\lambda) (\lambda I - A) = (C_0 + C_1 \lambda + \cdots + C_{n-1} \lambda^{n-1}) (\lambda I - A) = q(\lambda) I \]
Then multiplying out the middle term, it follows that for all $|\lambda|$ sufficiently large,
\[ a_0 I + a_1 I \lambda + \cdots + I \lambda^n = C_0 \lambda + C_1 \lambda^2 + \cdots + C_{n-1} \lambda^n \]
\[ - [C_0 A + C_1 A \lambda + \cdots + C_{n-1} A \lambda^{n-1}] \]
\[ = -C_0 A + (C_0 - C_1 A) \lambda + (C_1 - C_2 A) \lambda^2 + \cdots + (C_{n-2} - C_{n-1} A) \lambda^{n-1} + C_{n-1} \lambda^n \]
Then, using Corollary 8.9.3, one can replace $\lambda$ on both sides with $A$. Then the right side is seen to
equal 0. Hence the left side, $q(A) I$ is also equal to 0.

Here is an interesting and significant application of block multiplication. In this theorem, $q_M(t)$
denotes the characteristic polynomial, $\det(tI - M)$. The zeros of this polynomial will be shown
later to be eigenvalues of the matrix $M$. First note that from block multiplication, for the following
block matrices consisting of square blocks of an appropriate size,
\[ \begin{pmatrix} A & 0 \\ B & C \end{pmatrix} = \begin{pmatrix} A & 0 \\ B & I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & C \end{pmatrix} \]
\[ \det \begin{pmatrix} A & 0 \\ B & C \end{pmatrix} = \det \begin{pmatrix} A & 0 \\ B & I \end{pmatrix} \det \begin{pmatrix} I & 0 \\ 0 & C \end{pmatrix} = \det(A) \det(C) \]
Theorem 8.9.5 Let $A$ be an $m \times n$ matrix and let $B$ be an $n \times m$ matrix for $m \leq n$. Then

$$q_{BA}(t) = t^{n-m}q_{AB}(t),$$

so the eigenvalues of $BA$ and $AB$ are the same including multiplicities except that $BA$ has $n - m$ extra zero eigenvalues. Here $q_A(t)$ denotes the characteristic polynomial of the matrix $A$.

Proof: Use block multiplication to write

$$
\begin{pmatrix}
AB & 0 \\
B & 0
\end{pmatrix}
\begin{pmatrix}
I & A \\
0 & I
\end{pmatrix} = 
\begin{pmatrix}
AB & ABA \\
B & BA
\end{pmatrix}
$$

$$
\begin{pmatrix}
I & A \\
0 & I
\end{pmatrix}
\begin{pmatrix}
0 & 0 \\
B & BA
\end{pmatrix} = 
\begin{pmatrix}
AB & ABA \\
B & BA
\end{pmatrix}.
$$

Therefore,

$$
\begin{pmatrix}
I & A \\
0 & I
\end{pmatrix}^{-1}
\begin{pmatrix}
AB & 0 \\
B & 0
\end{pmatrix}
\begin{pmatrix}
I & A \\
0 & I
\end{pmatrix} = 
\begin{pmatrix}
0 & 0 \\
B & BA
\end{pmatrix}
$$

Since the two matrices above are similar, it follows that

$$
\begin{pmatrix}
0_{m \times m} & 0 \\
B & BA
\end{pmatrix},
\begin{pmatrix}
AB & 0 \\
B & 0_{n \times n}
\end{pmatrix}
$$

have the same characteristic polynomials. Thus

$$
\det \begin{pmatrix} tI_{m \times m} & 0 \\ -B & tI - BA \end{pmatrix} = \det \begin{pmatrix} tI - AB & 0 \\ -B & tI_{n \times n} \end{pmatrix} \quad (8.13)
$$

Therefore,

$$
t^n \det (tI - BA) = t^n \det (tI - AB) \quad (8.14)
$$

and so $\det (tI - BA) = q_{BA}(t) = t^{n-m} \det (tI - AB) = t^{n-m}q_{AB}(t)$. 

8.10 Exercises

1. Let $m < n$ and let $A$ be an $m \times n$ matrix. Show that $A$ is not one to one. Hint: Consider the $n \times n$ matrix $A_1$ which is of the form

$$A_1 \equiv \begin{pmatrix} A \\ 0 \end{pmatrix}$$

where the 0 denotes an $(n - m) \times n$ matrix of zeros. Thus $\det A_1 = 0$ and so $A_1$ is not one to one. Now observe that $A_1x$ is the vector,

$$A_1x = \begin{pmatrix} Ax \\ 0 \end{pmatrix}$$

which equals zero if and only if $Ax = 0$. 

2. Let \( \mathbf{v}_1, \ldots, \mathbf{v}_n \) be vectors in \( \mathbb{F}^n \) and let \( M(\mathbf{v}_1, \ldots, \mathbf{v}_n) \) denote the matrix whose \( i^{th} \) column equals \( \mathbf{v}_i \). Define
\[
d(\mathbf{v}_1, \ldots, \mathbf{v}_n) \equiv \det(M(\mathbf{v}_1, \ldots, \mathbf{v}_n)).
\]
Prove that \( d \) is linear in each variable, (multilinear), that
\[
d(\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_n) = -d(\mathbf{v}_1, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_n),
\]
and
\[
d(\mathbf{e}_1, \ldots, \mathbf{e}_n) = 1 \quad (8.16)
\]
where here \( \mathbf{e}_j \) is the vector in \( \mathbb{F}^n \) which has a zero in every position except the \( j^{th} \) position in which it has a one.

3. If \( A, B \) are similar matrices, show that they have the same determinant. Also show that they have the same characteristic polynomial.

4. Suppose \( f : \mathbb{F}^n \times \cdots \times \mathbb{F}^n \to \mathbb{F} \) satisfies \( \mathbf{S}\mathbf{M} \) and \( \mathbf{S}\mathbf{T} \) and is linear in each variable. Show that \( f = d \).

5. Show that if you replace a row (column) of an \( n \times n \) matrix \( A \) with itself added to some multiple of another row (column) then the new matrix has the same determinant as the original one. It was done in the chapter but go over it yourself.

6. Use the result of Problem 5 to evaluate by hand the determinant
\[
\det \begin{pmatrix}
1 & 2 & 3 & 2 \\
-6 & 3 & 2 & 3 \\
5 & 2 & 2 & 3 \\
3 & 4 & 6 & 4
\end{pmatrix}.
\]

7. Find the inverse if it exists of the matrix
\[
\begin{pmatrix}
e^t & \cos t & \sin t \\
e^t & -\sin t & \cos t \\
e^t & -\cos t & -\sin t
\end{pmatrix}.
\]

8. Let \( Ly = y^{(n)} + a_{n-1}(x)y^{(n-1)} + \cdots + a_1(x)y' + a_0(x)y \) where the \( a_i \) are given continuous functions defined on an interval, \( (a, b) \) and \( y \) is some function which has \( n \) derivatives so it makes sense to write \( Ly \). Suppose \( Ly_k = 0 \) for \( k = 1, 2, \ldots, n \). The Wronskian of these functions, \( y_i \) is defined as
\[
W(y_1, \ldots, y_n)(x) \equiv \det \begin{pmatrix}
y_1(x) & \cdots & y_n(x) \\
y'_1(x) & \cdots & y'_n(x) \\
\vdots & \ddots & \vdots \\
y^{(n-1)}_1(x) & \cdots & y^{(n-1)}_n(x)
\end{pmatrix}
\]
Show that for \( W(x) = W(y_1, \ldots, y_n)(x) \) to save space,
\[
W'(x) = \det \begin{pmatrix}
y_1(x) & \cdots & y_n(x) \\
\vdots & \ddots & \vdots \\
y^{(n-2)}_1(x) & \cdots & y^{(n-2)}_n(x) \\
y^{(n)}_1(x) & \cdots & y^{(n)}_n(x)
\end{pmatrix}.
\]
8.10. EXERCISES

Now use the differential equation, \( Ly = 0 \) which is satisfied by each of these functions, \( y_i \) and properties of determinants presented above to verify that \( W' + a_{n-1} (x) W = 0 \). Give an explicit solution of this linear differential equation, Abel’s formula, and use your answer to verify that the Wronskian of these solutions to the equation, \( Ly = 0 \) either vanishes identically on \((a, b)\) or never.

9. Show that the identity matrix is not similar to any other matrix.

10. Two \( n \times n \) matrices, \( A \) and \( B \), are similar if \( B = S^{-1} A S \) for some invertible \( n \times n \) matrix \( S \). Prove a theorem which is illustrated by the following picture.

\[
\begin{array}{|c|c|}
\hline
\text{same trace, characteristic polynomial, determinant} & \text{similar} \\
\hline
\end{array}
\]

Give an example of two matrices which are not similar but they have the same trace, characteristic polynomial and determinant.

11. Suppose the characteristic polynomial of an \( n \times n \) matrix \( A \) is of the form
\[
t^n + a_{n-1} t^{n-1} + \cdots + a_1 t + a_0
\]
and that \( a_0 \neq 0 \). Find a formula \( A^{-1} \) in terms of powers of the matrix \( A \). Show that \( A^{-1} \) exists if and only if \( a_0 \neq 0 \). In fact, show that \( a_0 = (-1)^n \det(A) \).

12. Let \( p(t) \) denote the characteristic polynomial of \( A \), show that \( p_c(t) = p(t - \varepsilon) \) is the characteristic polynomial of \( A + \varepsilon I \). Then show that if \( \det(A) = 0 \), it follows that \( \det(A + \varepsilon I) \neq 0 \) whenever \( |\varepsilon| \) is sufficiently small.

13. In constitutive modeling of the stress and strain tensors, one sometimes considers sums of the form \( \sum_{k=0}^{\infty} a_k A^k \) where \( A \) is a \( 3 \times 3 \) matrix. Show using the Cayley Hamilton theorem that if such a thing makes any sense, you can always obtain it as a finite sum having no more than \( n \) terms.

14. Recall you can find the determinant from expanding along the \( j^{th} \) column.
\[
\det(A) = \sum_i A_{ij} (\text{cof}(A))_{ij}
\]
Think of \( \det(A) \) as a function of the entries, \( A_{ij} \). Explain why the \( ij^{th} \) cofactor is really just
\[
\frac{\partial \det(A)}{\partial A_{ij}}.
\]

15. Let \( U \) be an open set in \( \mathbb{R}^n \) and let \( \mathbf{g} : U \to \mathbb{R}^n \) be such that all the first partial derivatives of all components of \( \mathbf{g} \) exist and are continuous. Under these conditions form the matrix \( D\mathbf{g} (\mathbf{x}) \) given by
\[
D\mathbf{g} (\mathbf{x})_{ij} \equiv \frac{\partial g_i (\mathbf{x})}{\partial x_j} \equiv g_{i,j} (\mathbf{x})
\]
The best kept secret in calculus courses is that the linear transformation determined by this matrix \( D\mathbf{g} (\mathbf{x}) \) is called the derivative of \( \mathbf{g} \) and it is the correct generalization of the concept of derivative of a function of one variable. Suppose the second partial derivatives also exist and are continuous. Then show that \( \sum_j (\text{cof}(D\mathbf{g}))_{ij} = 0 \). \textbf{Hint:} First explain why \( \sum_j g_{i,k} \text{cof} (D\mathbf{g})_{ij} = \delta_{jk} \det(D\mathbf{g}) \). Next differentiate with respect to \( x_j \) and sum on \( j \) using the equality of mixed partial derivatives. Assume \( \det(D\mathbf{g}) \neq 0 \) to prove the identity in this special case. Then explain using Problem 22 why there exists a sequence \( x_k \to 0 \) such that for \( \mathbf{g}_{x_k} (\mathbf{x}) = \mathbf{g} (\mathbf{x}) + x_k \mathbf{x} \), \( \det(D\mathbf{g}_{x_k}) \neq 0 \) and so the identity holds for \( \mathbf{g}_{x_k} \). Then take a limit to get the desired result in general. This is an extremely important identity which has surprising implications. One can build degree theory on it for example. It also leads to simple proofs of the Brouwer fixed point theorem from topology.
16. A determinant of the form
\[
\begin{vmatrix}
1 & 1 & \cdots & 1 \\
1 & a_0 & a_1 & \cdots & a_n \\
a_0^2 & a_1^2 & \cdots & a_n^2 \\
\vdots & \vdots & \ddots & \vdots \\
a_0^{n-1} & a_1^{n-1} & \cdots & a_n^{n-1} \\
a_0^n & a_1^n & \cdots & a_n^n
\end{vmatrix}
\]
is called a Vandermonde determinant. Show it equals \( \prod_{0 \leq i < j \leq n} (a_j - a_i) \). By this means to take the product of all terms of the form \((a_j - a_i)\) such that \(j > i\). **Hint:** Show it works if \(n = 1\) so you are looking at
\[
\begin{vmatrix}
1 & 1 \\
1 & a_0 & a_1
\end{vmatrix}
\]. Then suppose it holds for \(n - 1\) and consider the case \(n\). Consider the polynomial in \(t, p(t)\) which is obtained from the above by replacing the last column with the column \(\begin{pmatrix}1 & t & \cdots & t^n\end{pmatrix}^T\). Explain why \(p(a_j) = 0\) for \(i = 0, \ldots, n - 1\). Explain why \(p(t) = c \prod_{i=0}^{n-1} (t - a_i)\). Of course \(c\) is the coefficient of \(t^n\). Find this coefficient from the above description of \(p(t)\) and the induction hypothesis. Then plug in \(t = a_n\) and observe you have the formula valid for \(n\).

17. The example in this exercise was shown to me by Marc van Leeuwen and it helped to correct a misleading proof of the Cayley-Hamilton theorem presented in this chapter. If \(p(\lambda) = q(\lambda)\) for all \(\lambda\) or for all \(\lambda\) large enough where \(p(\lambda), q(\lambda)\) are polynomials having matrix coefficients, then it is not necessarily the case that \(p(A) = q(A)\) for \(A\) a matrix of an appropriate size. The proof in question read as though it was using this incorrect argument. Let
\[
E_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad E_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}
\]
Show that for all \(\lambda, (\lambda I + E_1)(\lambda I + E_2) = (\lambda^2 + \lambda) I = (\lambda I + E_2)(\lambda I + E_1)\). However,
\[
(NI + E_1)(NI + E_2) \neq (NI + E_2)(NI + E_1).
\]
Explain why this can happen. In the proof of the Cayley-Hamilton theorem given in the chapter, show that the matrix \(A\) does commute with the matrices \(C_i\) in that argument. **Hint:** Multiply both sides out with \(N\) in place of \(\lambda\). Does \(N\) commute with \(E_i\)?

18. Explain why the proof of the Cayley-Hamilton theorem given in this chapter cannot possibly hold for arbitrary fields of scalars.

19. Suppose \(A\) is \(m \times n\) and \(B\) is \(n \times m\). Letting \(I\) be the identity of the appropriate size, is it the case that \(\det(I + AB) = \det(I + BA)\)? Explain why or why not.

20. Suppose \(A\) is a linear transformation and let the characteristic polynomial be
\[
\det(\lambda I - A) = \prod_{j=1}^{q} \phi_j(\lambda)^{n_j}
\]
where the \(\phi_j(\lambda)\) are irreducible. Explain using Corollary why the irreducible factors of the minimum polynomial are \(\phi_j(\lambda)\) and why the minimum polynomial is of the form \(\prod_{j=1}^{q} \phi_j(\lambda)^{r_j}\) where \(r_j \leq n_j\). You can use the Cayley Hamilton theorem if you like.

21. Use the existence of the Jordan canonical form for a linear transformation whose minimum polynomial factors completely to give a proof of the Cayley Hamilton theorem which is valid for any field of scalars. **Hint:** First assume the minimum polynomial factors completely into linear factors. In this case, note that the characteristic polynomial is of degree \(n\) and is the
product of \((\lambda - \mu)\) where \(\mu\) is an eigenvalue and listed according to algebraic multiplicity. However, if there are multiple blocks corresponding to some \(\mu\), then the minimum polynomial will have such terms but fewer of them. If the minimum polynomial does not split, consider a splitting field of the minimum polynomial. Then consider the minimum polynomial with respect to this larger field. How will the two minimum polynomials be related? The two characteristic polynomials will be exactly the same, being defined in terms of the determinant of \(\lambda - A\). Show the minimum polynomial always divides the characteristic polynomial for any field \(F\).

22. Let \(q(\lambda)\) be a polynomial and \(C\) its companion matrix. Show the characteristic and minimum polynomial of \(C\) are the same and both equal \(q(\lambda)\).

23. Use the existence of the rational canonical form to give a proof of the Cayley Hamilton theorem valid for any field, even fields like the integers \(\mod p\) for \(p\) a prime. The proof in this chapter on determinants was fine for fields like \(\mathbb{Q}\) or \(\mathbb{R}\) where you could let \(\lambda \to \infty\) but it is not clear the same result holds in general.

24. Show that to find the eigenvalues of a matrix, it suffices to consider the roots of the characteristic polynomial. Hint: Use Cayley Hamilton theorem. This gives another way to find eigenvalues.

25. Recall that a matrix was diagonalizable if it was similar to a diagonal matrix. Suppose you have a matrix \(A\) whose entries are in \(F\) and the characteristic polynomial is the same as the minimum polynomial but the characteristic polynomial of the matrix has a repeated root. Can you show that the matrix cannot be diagonalizable in any field containing \(F\)?

26. For \(W\) a subspace of \(V\), \(W\) is said to have a complementary subspace \(W'\) if \(W \oplus W' = V\). Suppose that both \(W, W'\) are invariant with respect to \(A \in \mathcal{L}(V, V)\). Show that for any polynomial \(f(\lambda)\), if \(f(A)x \in W\), then there exists \(w \in W\) such that \(f(A)x = f(A)w\). A subspace \(W\) is called \(A\) admissible if it is \(A\) invariant and the condition of this problem holds.

27. Return to Theorem \(\text{巡回}\) about the existence of a basis \(\beta = \{\beta_{x_1}, \cdots, \beta_{x_p}\}\) for \(V\) where \(A \in \mathcal{L}(V, V)\). Adapt the statement and proof to show that if \(W\) is \(A\) admissible, then it has a complementary subspace which is also \(A\) invariant. Hint:

The modified version of the theorem is: Suppose \(A \in \mathcal{L}(V, V)\) and the minimum polynomial of \(A\) is \(\phi(\lambda)^m\) where \(\phi(\lambda)\) is a monic irreducible polynomial. Also suppose that \(W\) is an \(A\) admissible subspace. Then there exists a basis for \(V\) which is of the form \(\beta = \{\beta_{x_1}, \cdots, \beta_{x_p}, v_1, \cdots, v_m\}\) where \(\{v_1, \cdots, v_m\}\) is a basis of \(W\). Thus \(\mathrm{span}(\beta_{x_1}, \cdots, \beta_{x_p})\) is the \(A\) invariant complementary subspace for \(W\). You may want to use the fact that \(\phi(A)(V) \cap W = \phi(A)(W)\) which follows easily because \(W\) is \(A\) admissible. Then use this fact to show that \(\phi(A)(W)\) is also \(A\) admissible.

28. When you have an Abelian group \(V\) and a commutative ring with unity \(K\) such that the usual vector space operations hold

\[
k(v_1 + v_2) = kv_1 + kv_2, \quad (k_1 + k_2)v = k_1v + k_2v
\]

\[
k_1(k_2v) = (k_1k_2)v, \quad 1v = v
\]

then you call this \(V\) a \(K\) module. Thus, it is just a vector space except you have a ring of scalars rather than a field of scalars. Now suppose \(K = \mathbb{Z}\) the integers and \(V = \mathbb{Z}_m\) where \(m\) is some positive integer. Then if \(k \in K\) and \(\hat{a} \in \mathbb{Z}_m\), you define \(k\hat{a}\) in the usual way. Just add \(\hat{a}\) to itself \(k\) times or if \(k\) is negative, you just add \((-\hat{a}) = \overline{m - a}\) to itself \(|k|\) times. Explain why this is a \(\mathbb{Z}\) module. More generally, explain why an arbitrary Abelian group is a \(\mathbb{Z}\) module. However, show that there is no linearly independent set of elements of \(\mathbb{Z}_m\) which spans \(\mathbb{Z}_m\), although it is certainly true that \(1\) spans \(\mathbb{Z}_m\). Thus, when you replace a field with a ring, you loose the theorem that gives you a linearly independent subset of a spanning set.
29. Now suppose you have $K$ a commutative ring with unity and consider $K^n$. Show that

$$\{e_1, \cdots, e_n\}$$

spans $K^n$ and that if you have $\{a_1, \cdots, a_m\}$ for $m < n$, then $\{a_1, \cdots, a_m\}$ does not span $K^n$.

**Hint:** If it does span, then explain why you could get the following

$$A_{n \times m} P_{m \times n} = \begin{pmatrix} a_1 & \cdots & a_m \end{pmatrix} P = I_{n \times n}$$

Then consider something like this:

$$\begin{pmatrix} A_{n \times m} & 0 \end{pmatrix} \begin{pmatrix} P_{m \times n} & I_{n \times n} \end{pmatrix} = I_{n \times n}$$

Now consider Theorem 8.4.4 which still works if the entries of the matrix are from a commutative ring with unity. Is $\{e_1, \cdots, e_n\}$ also linearly independent? By this is meant one of the definitions given earlier that if you have a linear combination of these vectors equal to 0, then all of the scalars are zero. Since the scalars only come from a ring, you can’t conclude that this is the same thing as saying that no vector is a linear combination of the others.
Chapter 9

Modules And Rings *

9.1 Integral Domains And The Ring Of Polynomials

In this chapter is a very short introduction to modules and rings leading to canonical forms. The only rings considered here will be commutative rings. It all applies to the ring of polynomials \( F[x] \), but it extends to more general situations and this is included for the sake of interest. However, as far as linear algebra is concerned, you can just regard all of these things as pertaining to the ring \( F[x] \). This is a huge subject with a correspondingly huge collection of terminology, but I am only trying to include a little more than those parts which tie in well to what has been presented earlier, especially canonical forms. To see a thorough treatment of this subject, you should read a good book on modern algebra.

Definition 9.1.1 A commutative ring \( R \) is called an integral domain if in addition, whenever \( x \neq 0 \) and \( y \neq 0 \), it follows that \( xy \neq 0 \). A subset \( I \) of a commutative ring \( R \) is called an ideal if whenever \( r \in R \) and \( a \in I \), if whenever \( a, b \in I \), so is \( a + b \) and if \( a \in I \), so is \( -a \). (\( I \) is a subgroup of the additive group of \( R \)) Informally, \( RI = I \). A principal ideal is an ideal of the form \( \{ra : r \in R\} \) informally \( Ra \), usually written as \( (a) \). A principal ideal domain, p.i.d. is an integral domain in which the only ideals are principal ideals. This will be denoted as \( D \) in what follows. A Euclidean domain is an integral domain in which there is a special function \( \delta \) defined on the nonzero elements of \( D \) having value in \( \{0, 1, 2, \cdots \} \) such that if \( a, b \) are nonzero elements of \( D \), then there exist \( q, r \in D \) such that \( a = bq + r \) where \( \delta(r) < \delta(b) \) or else \( r = 0 \). In other words, it is like the polynomials in \( F[x] \) for \( F \) a field.

Note that \( Ra = (a) \) is indeed an ideal because \( (r_1a + r_2a) = a(r_1 + r_2) \) and so \( I \) is closed with respect to sums. Also if \( ra \in Ra \), then \( r(-a) = -ra \in Ra \) also and is the additive inverse. Thus \( Ra \) is indeed a subgroup of the additive group on \( R \). If \( ra \in Ra \), then if \( r \in R \), \( \hat{r}(ra) = (\hat{r}r)a \in Ra \). Hence \( Ra \) is an ideal.

The example we have in mind is the following, although you can also see that the integers \( \mathbb{Z} \) is also a principal ideal domain. The following theorem was partly done in Lemma \[\text{Lemma} \] but here is a different proof.

Theorem 9.1.2 Let \( F \) be a field and consider \( F[x] \), the polynomials in \( x \). Then this is a principal ideal domain. (Recall that two polynomials are equal if and only if all their coefficients coincide, this being the definition of what it means for equality to take place.)

Proof: It is obvious that \( F[x] \) is a commutative ring. Why is it an integral domain? Say

\[
p(x)q(x) \equiv \left( \sum_{i=0}^{m} a_ix^i \right) \left( \sum_{j=0}^{n} b_jx^j \right) = 0
\]
and suppose both \( p(x) \neq 0 \). Let \( a_r \) be the first nonzero coefficient of \( p(x) \) and let \( b_s \) be the first nonzero coefficient of \( q(x) \). Then one of the terms which must be zero is the coefficient of \( x^{r+s} \) which is

\[
\sum_{i+j=r+s} a_ib_jx^{r+s} = 0
\]

The only nonzero terms in the sum involve \( i \geq r \) and \( j \geq s \). However, \( i+j = r+s \) and so there is only one nonzero term in the sum and it is \( a_rb_s \) which must be zero, a contradiction to \( a_r, b_s \) both nonzero. Therefore, one of \( p(x) \) or \( q(x) \) must equal zero. Thus \( \mathbb{F}[x] \) is indeed an integral domain.

Next consider the question whether it is a principal ideal domain. Let \( I \) be an ideal. Suppose that it is not the zero ideal which is obviously a principal ideal but uninteresting. Then let \( g(x) \in I \) be a polynomial whose degree is as small as possible out of all non zero polynomials in \( I \). Then if \( f(x) \in I \), the division theorem implies that

\[
f(x) = g(x)q(x) + r(x)
\]

where \( r(x) \) is zero or else it has smaller degree than \( g(x) \). However,

\[
r(x) = f(x) - g(x)q(x)
\]

The right side is in \( I \) because it is a sum of two things in \( I \). Now this is a contradiction if \( r(x) \neq 0 \) because then it would have degree less than the smallest possible degree. It follows that \( r(x) = 0 \) and so \( f(x) = g(x)q(x) \) showing that everything in \( I \) is a multiple of \( g(x) \). Thus \( I \) is a principal ideal. \( \square \)

Note that the first part of the proof works if the coefficients of polynomials are only known to be in an integral domain. However, when we do division, we like to have coefficients in a field, but you might consider whether one could generalize to the situation in which the coefficients are only in an integral domain for the second part also.

In what follows the symbol \( \mid \) indicates “divides”. Thus \( a/b \) means \( b = ax \) for some \( x \).

**Observation 9.1.3** Note that, just as with matrices, if \( xy = yx = 1 \) and \( xz = zx = 1 \), then \( y = z \) so the inverse, if it exists, is unique. However, this is easier here because multiplication is commutative. In case \( D = \mathbb{F}[x] \), the invertibles are just non-zero scalars. \( x \) in \( D \) is invertible means there exists \( x^{-1} \in D \) such that \( xx^{-1} = 1 \). To see this last claim, If \( a(x) \) is invertible, then for some \( b(x) \),

\[
a(x)b(x) = 1
\]

but the degrees add and so the degree of both \( a(x) \) and \( b(x) \) must be 0. Hence \( a(x) \) is a nonzero scalar.

**Definition 9.1.4** Let \( D \) be an integral domain. Then \( p \in D \) is said to be prime if it is not zero and divided only by invertible elements of \( D \) and elements of the form \( xp \) where \( x \) is invertible. If \( a, b \in D \) then a greatest common divisor \( d \) denoted \( g.c.d \) satisfies \( d/a, d/b \) and if \( c/a, c/b \), then \( c/d \). If \( Da + Db = D1 = D \), then we say that \( a, b \) are relatively prime.

Note that invertible elements are prime and also if \( p \) is prime, so is \( xp \) where \( x \) is invertible. Here is why: If \( y/xp \) where \( x \) invertible and \( p \) prime, then \( xp = yk \) and so \( p = x^{-1}yk \) showing that \( x^{-1}y/p \) and so, since \( p \) is prime, either \( x^{-1}y = p \) or \( x^{-1}y = zp \) where \( z \) is invertible. In the second case, \( y = xp \) and so \( y \) is an invertible times \( p \). In the first case, \( y = xp \) so \( y \) is an invertible times \( p \). Thus \( xp \) is prime. As to invertibles being primes, if \( x \) is invertible, then if \( y/x \), then \( x = yz \) and so \( 1 = yx^{-1}z \) showing that \( y \) is invertible. Thus if \( y/x \), then \( y \) is invertible. Thus the only divisors of \( x \) are invertibles times \( x \) or invertibles. Thus \( x \) qualifies to be called prime.

When \( p = xq \) for \( x \) invertible, \( p, q \) are said to be associates or associated.

**Lemma 9.1.5** In a p.i.d. every pair of nonzero \( a, b \) has a greatest common divisor \( d \) and it can be written in the form

\[
d = sa + tb
\]

for some \( s, t \in D \). Also \( a, b \) are relatively prime if and only if there exists \( s, t \) such that

\[
1 = sa + tb
\]
9.1. INTEGRAL DOMAINS AND THE RING OF POLYNOMIALS

Proof: Consider the set $Da + Db$, all linear combinations of $a, b$. This is clearly an ideal and since this is a principal ideal domain, there exists $d$ such that $Dd = Da + Db$. Thus there exists $s, t$ such that $d = sa + tb$. Also, there exists $k$ such that $kd = 1a + 0b$ which shows that $d/a$. Similarly $d/b$. Now suppose $c/a$ and $c/b$ so $cx = a, cy = b$. Then $d = sx + ty = (sx + ty)c$ showing that $c/d$.

Next suppose $a, b$ are relatively prime. This means $Da + Db = D$. In particular, you can get $s, t$ such that $sa + tb = 1$ since $1 \in D$. Conversely, if for some $s, t$ you have $1 = sx + ty$, then it is obvious that $D = Da + Db$. ■

Here is a proposition which gives a condition for $a, b$ to be relatively prime.

Proposition 9.1.6 In a p.i.d. $a, b$ are relatively prime if and only if the only divisors of $a, b$ are the invertible elements of $D$.

Proof: If $a, b$ are relatively prime, then

$$1 = at + bs$$

for some $s, t \in D$. If $d$ divides both $a, b$, then $a = dx, b = dy$ for some $x, y$ and so

$$1 = at + bs = dxt + dys = d( xt + ys)$$

and so $d^{-1} = (xt + ys)$ and so if these are relatively prime then the only elements of $D$ which divide both are invertible elements of $D$.

Conversely suppose the only divisors are invertible. Consider the ideal $Da + Db$. It equals $Dd$ for some $d$ because this is a p.i.d. As above, $d/a, d/b$. By assumption, $d^{-1}$ exists and so $Dd = D$. Hence there exists $s, t$ such that $1 = sa + tb$ and so $a, b$ are relatively prime by Lemma 9.1.5. ■

Corollary 9.1.7 Let $p$ be a prime in a principal ideal domain $D$. If $p|ab$ then $p/a$ or $p/b$.

Proof: Say $p|ab$. Suppose that $p$ does not divide $a$. Since $p$ is prime, this means that if $k/p$ then $k$ is an invertible or an invertible times $p$. It follows that the only things from $D$ which divide both $p$ and $a$ are invertibles times $p$ or invertibles. However, the former does not occur because it is assumed that $p$ does not divide $a$. Thus $p, a$ are relatively prime because the only element of $D$ dividing them both is an invertible. This implies the greatest common divisor is invertible. Hence there are $s, t \in D$ such that

$$1 = sa + tp$$

which implies

$$b = sab + tpb$$

Now we use that $p|ab$ to write $ab = pz$. Then

$$b = spz + tpb = p(sz + tb)$$

showing that $p/b$. ■

What are examples of prime elements of our favorite integral domain $F[x]$?

Definition 9.1.8 In $F[x]$, we will say that $\phi(x)$ is irreducible if the only divisors of $\phi(x)$ are nonzero scalar multiples of $\phi(x)$ and scalars. That is, $\phi(x)$ is irreducible means that if $\psi(x)$ divides $\phi(x)$, then $\psi(x)$ is a nonzero scalar or a nonzero scalar multiple of $\phi(x)$.

As mentioned above, in $F[x]$, the invertible elements are nonzero scalars.

The above definition just gives the content of the following lemma. It is just a restatement of the above definition of what it means to be prime.

Lemma 9.1.9 In $F[x], \phi(x)$ is prime if and only if it is irreducible.
As discussed much earlier, every polynomial in \( F[x] \) can be factored as a product of prime (irreducible) elements of \( F[x] \). What about an arbitrary p.i.d.? It turns out this is true in this setting also and this is very interesting. First we need the following wonderful lemma. It says that there is an ascending chain condition on the ideals. Such rings which have an ascending chain condition are called Noetherian rings.

**Lemma 9.1.10** Let \( D \) be a p.i.d. (principal ideal domain). Also let \( I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots \) be an ascending chain of ideals. Then eventually, \( I_n = I_{n+1} = I_{n+1} = \cdots \) so the \( I_k \) do not change as \( k \) increases.

**Proof:** Let \( I = \cup I_i \). This is a subgroup of the additive group of \( D \) because if \( a \in I \), then eventually \( a \in I_n \) and so \(-a \in I_n \subseteq I \). Similarly, if \( a, b \in I \), then for large enough \( n \), both are in \( I_n \) and so their sum is in \( I_n \subseteq I \). If \( d \in D \) and \( a \in I \), then \( d \in I_m \) for some \( m \) and so \( da \in I_m \subseteq I \). So yes, \( I \) is an ideal. Therefore, \( I = Da = (a) \) for some \( a \) because this is a principal ideal domain. Then \( a \in I \) and so \( a \in I_k \) for all \( k \geq n \) for some \( n \). Hence \( I = Da \equiv (a) \subseteq I_k \) for all \( k \) sufficiently large. ■

Recall again the definition of a prime. \( p \in D \) is said to be prime if it is divided only by invertible elements of \( D \) and elements of the form \( xp \) where \( x \) is invertible.

What follows is the prime factorization theorem analogous to the fundamental theorem of arithmetic. Note the following example in \( \mathbb{Q}[x] \),

\[
(x^2 + 2) \left( x^2 + x + 1 \right) = (2x^2 + 4) \left( \frac{1}{2} x^2 + \frac{1}{2} x \right) + \frac{1}{2}
\]

both on the right are irreducible over \( \mathbb{Q} \) and so you can’t expect uniqueness in such a prime factorization for a p.i.d. What you can expect is uniqueness up to associates.

**Theorem 9.1.11** Let \( D \) be a p.i.d. and let \( a \in D, a \neq 0 \) and \( a \) is not invertible.\(^1\) Then there are primes \( p_1, \cdots, p_m \) none invertible such that \( a = \prod_{i=1}^{m} p_i \). If also \( a = \prod_{j=1}^{n} q_j \), where each \( q_j \) is a noninvertible prime, then \( m = n \) and there is a permutation such that \( x_k p_k = q_{i_k} \) for some invertible \( x \). (The two primes \( p_k, q_{i_k} \) are said to be “associate” when this happens, or they are said to be “associated”.)

**Proof:** Say \( a \) is not invertible and suppose it has no prime factorization. Then \( a \) is itself not prime. Thus \( a \) has a divisor \( a_1 \) which is not invertible and also is not an invertible multiple of \( a \). Thus \( a = a_1 b_1 \). If \( b_1 \) were invertible, then \( a_1 = b_1^{-1} a \) and this violates what is known about \( a_1 \) that it is not an invertible times \( a \). Therefore, \( b_1 \) is not invertible. If \( b_1 = xa \), then you would have \( a = a_1 xa \) and so, \( a (1 - xa) = 0 \). Since \( a \neq 0 \), \( 1 = a_1 x \) which would mean that \( a_1 \) is invertible which it is not. Therefore, neither \( a_1 \) nor \( b_1 \) are invertible and neither are invertible multiples of \( a \).

If these each had a prime factorization consisting of non invertible primes, then so would \( a \equiv a_0 \). Let \( a_1 \) be the one which does not have such a prime factorization. Then repeating the above argument for \( a_1 \), it follows that \( a_1 = a_2 b_2 \) where neither factor is invertible nor an invertible multiple of \( a_1 \) and as just explained, one fails to have a prime factorization. Let \( a_2 \) be the one which does not have the prime factorization consisting of non invertible primes. Then continuing this way, you get a sequence

\[
a_0, a_1, a_2, a_3, \cdots
\]

where \( a_n \) is a non invertible multiple of \( a_{n+1} \). It follows that the ideals \( (a_0), (a_1), (a_2), \cdots \) form an ascending chain which never becomes constant because \( a_{n+1}/a_n \), which is a contradiction to Lemma 9.1.10.

To see why these are strictly increasing, note first that \( (a_n) \subseteq (a_{n+1}) \). If they were equal, then \( a_{n+1} = ba_n \). However, from the construction, \( a_n = ba_{n+1} \) and so \( a_n = bba_n \). Thus \( a_n \left( 1 - bb \right) = 0 \), and since this is an integral domain, it follows that \( 1 - bb = 0 \) so \( b \) is invertible contrary to the construction. Thus strict inequality holds.

\(^1\)In our favorite example of \( F[x] \), the invertible elements are just the nonzero scalars.
If \( a^{-1} \) were to exist, this argument would fail because you would have \( 1 = a q^{-1} a_0 \in Da_0 \) and so this ideal would be all of \( D \) and there would be no contradiction because the ideals would all be the same.

This takes care of the existence of the factorization.

Now suppose \( \prod_{i=1}^m p_i = a \). If \( m = 1 \), then \( a \) is a prime and so if \( a = p = q_1 \cdots q_n \) for \( n > 1 \) with each \( q_i \) a noninvertible prime, each \( q_i/p \) and so each must either be invertible or an invertible multiple of \( p \). Since, by assumption, none are invertible, \( p = a q_1 \) and so \( p = a^{-1} p q_2 \cdots q_n \). Similarly, \( q_2 b = p \) and so \( q_2 = b^{-1} p \). Then you get

\[
p = (ba)^{-1} p q_3 \cdots q_n
\]

and cancelling \( p \), it follows that \( p \) is invertible which it is not. The error occurred in assuming there was more than one factor in the \( q_i \). Hence in this case that \( m = 1 \), the second claim about uniqueness follows. Assume this claim is true for \( m - 1, m > 1 \) that if one of \( k, l \) is no larger than \( m - 1 \), and if

\[
\prod_{i=1}^k p_i = \prod_{j=1}^l q_j,
\]

then \( k = l \) and there is a permutation such that \( q_j = p_i x \) for some invertible \( x \).

Then if \( \prod_{i=1}^m p_i = \prod_{j=1}^n q_j \) where the primes are not invertible, we can assume \( n > m - 1 \) because if not, the induction hypothesis would apply to draw the desired conclusion. Then \( q_n \) divides some \( p_i \). Assume it divides \( p_m \) for the sake of simplicity. Then \( q_n \) is not invertible and so it must be of the form \( q_n = xp_m \) where \( x \) is invertible. Thus

\[
p_m \prod_{i=1}^{m-1} p_i = x \prod_{j=1}^{n-1} q_j p_m
\]

Thus, \( p_m \left( \prod_{i=1}^{m-1} p_i - x \prod_{j=1}^{n-1} q_j \right) = 0 \)

and now by induction, \( m - 1 = n - 1 \) and there is a permutation of the \( q_j \) such that \( q_{j_i} = x_i p_i \) where \( x_i \) is invertible. Recall that if \( q \) is prime, so is \( x q \) for \( x \) an invertible.

The next proposition is like Corollary 9.1.11.

**Proposition 9.1.12** Let \( a/b \) where \( b = \prod_{i=1}^n p_i^{k_i} \) is not invertible. Then \( a \) is either invertible or \( a = x \prod_{i=1}^n p_i^{m_i} \) where \( x \) is an invertible and \( m_i \leq k_i \), the \( p_i \) being distinct non-invertible primes.

**Proof:** We have \( b = ak \). Then by Theorem 9.1.11 assuming that \( a \) is not invertible,

\[
b = \prod_{i=1}^n p_i^{k_i} = \prod_{i=1}^n q_i^{m_i}, \quad k = \prod_{i=1}^s q_i^{m_i}, \quad \prod_{j=1}^t q_j^{m_j}
\]

where the \( q \) are non-invertible primes and by collecting the \( q, \hat{q} \) primes, we can assume all of the primes in the right side are unique. By the uniqueness part of Theorem 9.1.11, each \( q_i \) and \( \hat{q}_j \) is an invertible times one of the \( p_i \) and there are enough \( p_i \) to use up all of the \( q \) on the right. Thus, in particular, \( a = \prod_{i=1}^s q_i^{m_i} = x \prod_{i=1}^s p_i^{m_i} \) where \( m_i \leq k_i, s \leq n \), and \( x \) is some invertible obtained as a product of invertibles.
9.2 Modules And Decomposition Into Cyclic Sub-Modules

Definition 9.2.1 Let \( R \) be a commutative ring. Then \( M \) is an \( R \) module\(^2\) if it acts just like a vector space except for having the coefficients come from \( R \) rather than a field. Thus

1. \( M \) is an Abelian group, and for \( r, r_1, r_2 \in R \) and \( m, n \in M \),
2. \( (r_1 + r_2)m = r_1m + r_2m \)
3. \( r(m + n) = rm + rn \)
4. \( r_1(r_2m) = (r_1r_2)m \)
5. \( 1m = m \)

Then, just as in the case of vector spaces, \( 0 \) in \( M \) is unique. So is \(-m\). These assertions only use the fact that \( M \) is an Abelian group so there is really no change here. In addition, \( 0m = 0 \). This follows from

\[
0m = (0 + 0)m = 0m + 0m
\]

Now add \(-0m\) to both sides. Also \((-1)m = -m\) where \(-1\) is the additive inverse of 1 in \( R \). This follows from

\[
0 = 0m = (1 + (-1))m = m + (-1)m
\]

\[
-m = (-1)m
\]

All this is just like the case of a vector space.

Here we will only consider the case where the ring is a p.i.d. and it will be written as \( D \). Modules will be written in capital letters like \( A, B, M \), etc. Elements of \( D \) will usually be written as Greek letters except for primes with will be denoted as \( p \).

Definition 9.2.2 Let the ring \( D \) be a p.i.d., a principal ideal domain and let \( M \) be a module as discussed above. Then \( M \) is called a torsion module if for every \( m \in M \), there exists \( \alpha \in D, \alpha \neq 0 \), such that \( \alpha m = 0 \). The set of all such \( \alpha \) is called \( \text{ann}(m) \). This is short for annihilators of \( m \). The module \( M \) is said to be finitely generated if \( M = Da_1 + \cdots + Da_p \) for some \( a_i \in M \). This is like saying that \( M = \text{span}(a_1, \cdots, a_p) \) so this notation will be used when convenient. It isn’t quite the same because \( D \) is not a field, but it denotes the same kind of thing where scalars, this time from a ring are multiplied by elements of an Abelian group \( M \) and added together. The submodule \( Da \) for \( a \in M \) is called a cyclic sub module. More generally, \( N \) is a submodule of a module \( M \) if it is a subgroup and is itself a module.

What is an example of a finitely generated torsion module? The integral domain will be \( \mathbb{F}[x] \) and the module will be a finite dimensional vector space \( V \). This example will also involve a given \( L \in \mathbb{L}(V,V) \). Then it is necessary to define multiplication of something in the integral domain with something in \( V \). Here it is.

\[
g(x)v \equiv g(T)v
\]

In particular, if \( g(x) \) is a constant, \( \rho \) then \( g(x)v = \rho v \). Thus \( V \) is clearly a torsion module because if \( p(x) \) is the minimum polynomial of \( L \), then \( p(x) \in \text{ann}(v) \) for every \( v \in V \). Is this a finitely generated module? This seems to be pretty easy. Let \( \{v_1, \cdots, v_n\} \) be a basis. Then a typical \( v \) is of the form

\[
v = \alpha_1v_1 + \cdots + \alpha_nv_n \in Dv_1 + \cdots + Dv_n.
\]

If \( g(x) \in \mathbb{F}[x] \),

\[
g(x) = p(x)q(x) + r(x)
\]

where the degree of \( r(x) \) is less than the degree of \( p(x) \). Thus

\[
g(x)v = r(T)v
\]

This observation will be important later. First is an interesting proposition.

\(^2\)This is a left \( R \) module.
9.3. A DIRECT SUM DECOMPOSITION

Proposition 9.2.3 \( \text{ann} (m) \) is an ideal.

Proof: It is a subgroup of the additive group of \( D \) because if \( \alpha, \beta \in \text{ann} (m) \),

\[
(\alpha + \beta) (m) = \alpha (m) + \beta (m) = 0
\]

Also, if \( \alpha \in \text{ann} (m) \), then \( -\alpha (m) = (-1) \alpha (m) = 0 \). If \( \alpha \in \text{ann} (m) \), then \( \beta \alpha (m) = \beta (\alpha m) = 0 \) so it is indeed an ideal.

Note that if the ring is a principal ideal domain, this implies right away that \( \text{ann} (m) = (\alpha) \) for some \( \alpha \in D \).

Definition 9.2.4 When \( M_i \) is a submodule of \( M \) we write

\[
M_1 + \cdots + M_m = M_1 \oplus \cdots \oplus M_n
\]

when for each \( i \), \( M_i \cap \sum_{j \neq i} M_j = 0 \). This is equivalent to the following: If \( \sum_i m_i = 0 \) for \( m_i \in M_i \), then each \( m_i = 0 \). This is called a direct sum.

9.3 A Direct Sum Decomposition

As in [20], for \( p \) a noninvertible prime, define

\[
M_p \equiv \{ m \in M : p^k m = 0 \text{ for some } k \in \mathbb{N} \}
\]

That is, eventually \( p^k m = 0 \). Right now, it might be possible that \( k \) could change for different \( m \in M \). Note that if \( p \) is invertible, then \( M_p = 0 \) so nothing is lost from considering only non-invertible primes. It is obvious that \( M_p \) is a subgroup of the module \( M \) and is itself a module.

Proposition 9.3.1 Let \( p_1, \ldots, p_n \) be distinct primes (not associates), none of which is invertible. Let \( M \) be a module (torsion or not) over a p.i.d. \( D \). Then

\[
(M_{p_1} + \cdots + M_{p_{j-1}} + M_{p_{j+1}} + \cdots + M_{p_n}) \cap M_{p_j} = 0
\]

Proof: This follows from the observation that \( \prod_{i \neq j} p_i^{k_i} \) and \( p_j^{k_j} \) are relatively prime. If \( q \) divides the second, then by Proposition 9.2.1 it is of the form \( \xi p_j^{m_j}, m_j \leq k_j \) where \( \xi \) is an invertible. If \( q \) divides the first, then by the same proposition, it is \( \zeta \prod_{i \neq j} p_i^{m_i}, m_i \leq k_i \) for an invertible \( \zeta \). Since the primes are distinct, we must have all \( m_j, m_i \) equal to 0 and \( q \) is some invertible \( \xi \) and so these two, \( \prod_{i \neq j} p_i^{k_i} \) and \( p_j^{k_j} \) are relatively prime as claimed. Therefore, there exist \( \sigma, \tau \) such that

\[
1 = \sigma \prod_{i \neq j} p_i^{k_i} + \tau p_j^{k_j}
\]

If \( m \in (M_{p_1} + \cdots + M_{p_{j-1}} + M_{p_{j+1}} + \cdots + M_{p_n}) \cap M_{p_j} \), then \( m = \sum_{i \neq j} m_i \) and so there exist \( k_i, k_j \) such that \( p_i^{k_i} m_i = 0 \) and \( p_j^{k_j} m = 0 \). Then do both sides of * to \( m \).

\[
m = \left( \sigma \prod_{i \neq j} p_i^{k_i} \right) \left( \sum_{i \neq j} m_i \right) + \tau p_j^{k_j} m = 0
\]

This yields \( m = 0 \) and verifies the conclusion of the proposition. □

It follows from Proposition 9.2.1 that if \( m_i \in M_{p_i} \), and if \( \sum m_i = 0 \), then each \( m_i = 0 \) because \( m_i \in M_{p_i} \cap (M_{p_1} + \cdots + M_{p_{j-1}} + M_{p_{j+1}} + \cdots + M_{p_n}) \). The converse is also true. If whenever \( \sum m_i = 0 \), \( m_i \in M_{p_i} \), then each \( m_i = 0 \), then if \( m \in M_{p_j} \cap (M_{p_1} + \cdots + M_{p_{j-1}} + M_{p_{j+1}} + \cdots + M_{p_n}) \), then \( m = \sum_{j \neq i} m_i \), \( m_i \in M_{p_i} \), and so \( 0 = -m - \sum_{j \neq i} m_i \) which implies \( m = 0 \) and incidentally, each \( m_i = 0 \).

Now suppose that \( M \) is a finitely generated torsion module for a p.i.d. Then the following lemma will be the basis for a decomposition theorem in terms of a direct sum of some \( M_{p_i} \) for \( p_i \) a non-invertible prime.
Lemma 9.3.2 Let $M$ be a non zero finitely generated ($M = Dz_1 + \cdots + Dz_n$) torsion module over a p.i.d. $D$. Then there exist primes $\{p_1, \ldots, p_n\}$ and corresponding positive integers $k_i$ such that $\text{ann}(M) = \left(\prod_{i=1}^{n} p_i^{k_i}\right)$.

**Proof:** By assumption, $M = Dz_1 + \cdots + Dz_n$. Since $M$ is a torsion module, $\text{ann}(z_i)$ is nonempty, equal to $(\alpha_i)$. Say $\alpha_i = \prod_{j=1}^{m_i} p_{ij}^{k_j}$ where the $p_{ij}$ are non-invertible primes. Then define

$$\sigma = \prod_{i=1}^{n} p_i^{k_i}$$

as the least common multiple of the $\alpha_i$. Let $I \equiv \left(\prod_{i=1}^{n} p_i^{k_i}\right)$. Then $I$ is an ideal and if $\alpha \in I$, then $\alpha z_i = 0$ for each $i$. Hence $\alpha m = 0$ for every $m \in M$. Thus $I \subseteq \text{ann}(M)$, the set of all $\beta$ such that $\beta m = 0$ for all $m$. Conversely, suppose $\beta \in \text{ann}(M)$. Then in particular $\beta z_i = 0$ and so $\alpha_i/\beta$ because $\beta \in \text{ann}(z_i)$. Thus $\beta$ is a multiple of all of the $\alpha_i$ and it follows that the least common multiple $\sigma$ must divide $\beta$. Hence $\beta \in (\sigma)$. ■

Next is a lemma about direct sums.

Lemma 9.3.3 Let $m \in M$ a module over a p.i.d. $D$, and let $\alpha m = 0$, for $\alpha = \prod_{i=1}^{n} p_i^{k_i}$ where the $p_i$ are non-invertible distinct primes. Let

$$\hat{M}_{p_i} = \left\{m \in M : p_i^{k_i} m = 0 \right\} \subseteq M_{p_i},$$

(Note that here $k_i$ is fixed.) Then

$$Dm \subseteq \hat{M}_{p_1} \oplus \cdots \oplus \hat{M}_{p_n} \subseteq M_{p_1} \oplus \cdots \oplus M_{p_n}$$

**Proof:** If $n = 1$, then this is clearly true. Suppose it is true for $n - 1 \geq 1$. Say $\alpha m = 0$ where

$$\alpha = \prod_{i=1}^{n} p_i^{k_i}$$

Let $M' \equiv \left\{k \in M : \prod_{i=1}^{n-1} p_i^{k_i} k = 0 \right\}$. Then this is a module contained in $M$. Also $p_n^{k_n} m \in M'$. Therefore, by induction,

$$Dp_n^{k_n} m \subseteq \hat{M}_{p_1} \oplus \cdots \oplus \hat{M}_{p_{n-1}}$$

(*)

where

$$\hat{M}_{p_i} \equiv \left\{a \in M' : p_i^{k_i} a = 0 \right\}$$

Now there are $\sigma, \tau$ such that

$$1 = \sigma p_n^{k_n} + \tau \prod_{i=1}^{n-1} p_i^{k_i}$$

(**)

Therefore, if $k \in M'$, do both sides of the above to $k$ to conclude that $k = \sigma p_n^{k_n} k$. Then from $*$,

$$p_n^{k_n} m \Rightarrow \sum_{i=1}^{n-1} k_i, \ k_i \in \hat{M}_{p_i} \subseteq \hat{M}_{p_i}$$

Now also each $k_i \in M'$ and so doing both sides of $**$ to $k_i$ gives $k_i = \sigma p_n^{k_n} k_i$. It follows that the sum on the right equals

$$\sum_{i=1}^{n-1} \sigma p_n^{k_n} k_i,$$

It follows that

$$p_n^{k_n} \left( m - \sum_{i=1}^{n-1} \sigma k_i \right) = 0$$
Therefore,

\[ m - \sum_{i=1}^{n-1} \sigma k_i = k_n \in \hat{M}_{p_n} \]

It follows that

\[ Dm \subseteq \hat{M}_{p_1} + \cdots + \hat{M}_{p_n} \subseteq M_{p_1} + \cdots + M_{p_n} \]

However, this last equals \( M_{p_1} \oplus \cdots \oplus M_{p_n} \) thanks to Proposition 9.3.1. Also,

\[ \hat{M}_{p_1} + \cdots + \hat{M}_{p_n} = \hat{M}_{p_1} \oplus \cdots \oplus \hat{M}_{p_n} \]

by this proposition because \( \hat{M}_{p_i} \subseteq M_{p_i} \). ■

The following is the main result.

**Theorem 9.3.4** Let \( M \) be a non zero torsion module for a p.i.d. \( D \) and suppose that

\[ M = Dz_1 + \cdots + Dz_p \]

so that it is finitely generated. Then there exist primes \( \{p_i\}_{i=1}^n \) distinct in the sense that none is an invertible multiple of another such that

\[ M = M_{p_1} \oplus \cdots \oplus M_{p_n} \]

For \( (\beta) = \text{ann}(M), \beta = \prod_{i=1}^n p_i^{k_i}, \) a product of non-invertible primes, it follows

\[ \hat{M}_{p_i} \equiv \left\{ m \in M : p_i^{k_i}m = 0 \right\} = \left\{ m \in M : p_i^{k_i}m = 0 \text{ for some } k \right\} = M_{p_i} \]

which is not dependent on the spanning set.

**Proof:** Let \( (\beta) = \text{ann}(M) \). It exists by Lemma 9.3.2 above. Thus \( \beta \) is not invertible. Also if \( (\alpha) = \text{ann}(M) \), one needs \( \alpha/\beta \) and \( \beta/\alpha \) so the two are associates meaning that one is an invertible times the other. Let \( \beta = \prod_{i=1}^n p_i^{k_i} \) where this is the prime factorization of \( \beta \) in terms of non-invertible primes. Note this is well defined up to associates. Let \( \hat{M}_{p_i} \equiv \left\{ m \in M : p_i^{k_i}m = 0 \right\} \). Then by Lemma 9.3.3,

\[ Dz_k \subseteq \hat{M}_{p_1} \oplus \cdots \oplus \hat{M}_{p_n} \]

Including this for each \( z_k \), it follows that

\[ M \subseteq \hat{M}_{p_1} + \cdots + \hat{M}_{p_n} \subseteq M_{p_1} + \cdots + M_{p_n} \subseteq M \]

and so all inclusions are equals. By Proposition 9.3.1,

\[ M_{p_1} + \cdots + M_{p_n} = M_{p_1} \oplus \cdots \oplus M_{p_n} \]

and so

\[ M = M_{p_1} \oplus \cdots \oplus M_{p_n} \]

Now it follows that \( M_{p_i} = \hat{M}_{p_i} \). This is shown next. It is clear that \( \hat{M}_{p_i} \subseteq M_{p_i} \) from the definition. Let \( k \in M_{p_i} \). Then

\[ k = \sum_{j \neq i} k_j + k_i, \ k_s \in \hat{M}_{p_s} \subseteq M_{p_s} \]

Since \( M = M_{p_1} \oplus \cdots \oplus M_{p_n} \) it follows that \( k = k_i \in \hat{M}_{p_i} \), the other \( k_j \) equalling zero. ■

Note that this theorem is a lot like Theorem 6.1.10. It is more general but the ideas are very similar.
9.4 Quotients

One can consider quotients of modules. This involves a set of equivalence classes as described below.

**Definition 9.4.1** Let $A$ be a $D$ module and let $B$ be a submodule. Then $A/B$ denotes sets of the form $a + B$ defined by $\{a + b : b \in B\}$, with the operations defined by

\[
a + B + (\hat{a} + B) \equiv a + \hat{a} + B \\
\lambda (a + B) \equiv \lambda a + B
\]

Also, for $R$ a commutative ring and $I$ an ideal, one can define $R/I$ in the form $\alpha + I$ given by $\{\alpha + \lambda : \lambda \in I\}$. Then the operations are defined as above with $I$ taking the place of $B$. Also in the case of the ring and an ideal, define

\[
(\alpha + I)(\beta + I) \equiv \alpha \beta + I
\]

The main result about quotient spaces is in the following. It will be reminiscent of what was done with the field $\mathbb{Z}_p$ for a $p$ a prime.

**Proposition 9.4.2** $A/B$ is a $D$ module. Also $R/I$ is an $R$ module, this for an arbitrary commutative ring $R$. If $R$ is a p.i.d., then $R/(p)$ is a field if $p$ is a prime.

**Proof:** Note that $a + B = \hat{a} + B$ is the same as saying that $a - \hat{a} \in B$. Are the operations well defined? If $a + B = \hat{a} + B$ and $a_1 + B = a + B$, $\hat{a}_1 + B = \hat{a} + B$, does it follow that

\[
a + B + (\hat{a} + B) = a_1 + B + (\hat{a}_1 + B)
\]

Is

\[
a + \hat{a} + B = a_1 + \hat{a}_1 + B?
\]

Of course this is so because $a + \hat{a} = (a_1 + \hat{a}_1) \in B$ since $B$ is a group with respect to addition. Next suppose

\[
a + B = \hat{a} + B
\]

Is

\[
\lambda a + B = \lambda \hat{a} + B?
\]

This is also so because $B$ is a module so $\lambda a - \lambda \hat{a} = \lambda (a - \hat{a}) \in B$. The operations are indeed well defined and so this is indeed a $D$ module.

The claim about $R/I$ is entirely similar except here one needs to consider the operation of multiplication. Why is it well defined? Say $\alpha + I = \hat{a} + I$ and $\beta + I = \hat{\beta} + I$. Then $\alpha - \hat{a} \in I$ and $\beta - \hat{\beta} \in I$. This operation of multiplication is well defined if and only if $\alpha \beta - \hat{a} \hat{\beta}$ is in $I$. However, this difference is

\[
\alpha (\beta - \hat{\beta}) + \hat{\beta} (\alpha - \hat{a}) \in I
\]

because $I$ is an ideal. Thus the operation is well defined and $R/I$ is an $R$ module and in addition, it is also a ring.

Consider the last claim where $p$ is prime. If $\rho \in R$ and $\rho + (p) \neq 0$, this says nothing more than that $p$ fails to divide $\rho$. Thus the two are relatively prime and so it follows that there exist $\sigma, \tau \in R$ such that

\[
1 = \sigma \rho + \tau p
\]

Thus $\tau + (p)$ is the inverse of $\rho + (p)$ because the above says that

\[
(\tau + (p)) (\rho + (p)) \equiv \tau \rho + (p) = 1 - \sigma \rho + (p) = 1 + (p)
\]

which means that $(\tau + (p))^{-1} = (\rho + (p))$ with $1 + (p)$ being the multiplicative identity in $R/(p)$. ■
9.5 Cyclic Decomposition

First here is a definition.

**Definition 9.5.1** Let $M$ be a module. It is called a Noetherian module if whenever there is an increasing sequence of sub modules $\{N_k\}, \cdots, N_k \subseteq N_{k+1}, \cdots$, it follows that for large enough $k$ all the submodules are equal. (Ascending chains of submodules are eventually constant.) A mapping $\theta$ from a $R$ module $M$ to another $R$ module $N$ is a morphism if it preserves operations. That is,

$$\theta(m_1 + m_2) = \theta m_1 + \theta m_2, \quad \theta(\sigma m) = \sigma \theta(m)$$

It is just like “linear” except that here the scalars come from a ring.

When is a module $M$ Noetherian? A sufficient condition is for it to be a finitely generated module over a $p.i.d$. This is explained next.

**Lemma 9.5.2** Suppose

$$0 \rightarrow A \xrightarrow{\theta} B \xrightarrow{\eta} C \rightarrow 0$$

is a short exact sequence of modules for a $p.i.d$ $D$. Recall this means that these mappings are morphisms, $\eta \circ \theta = 0, \theta$ is one to one, $\eta$ is onto, and $\ker(\eta) = \theta(A)$. Then if $A, C$ are Noetherian, so is $B$.

**Proof:** Suppose $B_n$ is an ascending chain of submodules in $B$. Then $\eta(B_n)$ is an ascending chain of submodules in $C$ and so it is eventually constant. Also $\theta^{-1} B_n$ is an ascending chain of submodules in $A$. These are submodules because they are clearly additive subgroups and if $\alpha \in D, \alpha \theta^{-1}(B_n) = \theta^{-1}(\alpha B_n) = \theta^{-1}(B_n)$. Note that $\theta^{-1} B_n$ cannot be empty because it must contain $0$. Thus there exists $m$ such that if $n \geq m, \theta^{-1}(B_n) = \theta^{-1}(B_m) and \eta(B_n) = \eta(B_m)$. Let $b \in B_n$. It must be shown that $b \in B_m$. Since $\eta(B_n) = \eta(B_m)$, it follows that there is $\hat{b} \in B_m$ such that $\eta b = \eta \hat{b}$ so $(b - \hat{b}) \in \ker(\eta) = \theta(A)$ and so $\theta^{-1}(b - \hat{b}) \in \theta^{-1} B_n = \theta^{-1} B_m$. Hence $b - \hat{b} = c \in B_m$. Then $b = \hat{b} + c \in B_m$ showing that $B_n = B_m$ for $n \geq m$. ■

Now observe the following.

**Lemma 9.5.3** Suppose $A, B$ are modules and $\tau : A \rightarrow B$ is a morphism. Then if $C$ is a submodule of $B$, it follows that $\tau^{-1}(C)$ is a submodule of $A$ which contains $\ker(\tau)$.

**Proof:** If $a, b \in \tau^{-1}(C)$, then $\tau(a) + \tau(b) = \tau(a + b) \in C$ and so $a + b \in \tau^{-1}(C)$. It is similarly the case that if $a \in \tau^{-1}(C)$, then so is $-a$. Thus $\tau^{-1}(C)$ is a submodule of $A$. If $b \in \ker(\tau)$, then $\tau(b) = 0 \in C$ and so, $b \in \tau^{-1}(C)$.

Now note that a $p.i.d$. $D$ is a Noetherian module for $D$. This is because it has the ascending chain condition on ideals which are just the modules for $D$. Consider

$$0 \rightarrow D^{n-1} \xrightarrow{\theta} D^n \xrightarrow{\eta} D \rightarrow 0$$

where $\theta(b) \equiv (b,0), \eta(c) \equiv c_n$ where $c = (c_1, \cdots, c_n)$. Then this is a short exact sequence. It is clear that $\theta$ is one to one and that $\eta$ is onto. Also $\eta(\theta(b)) = \eta((b,0)) = 0$ and $\ker(\eta) = \{(b,0) : b \in D^{n-1}\} = \theta(D^{n-1})$.

Now consider the above lemma for $n = 2$ to conclude $D^2$ is Noetherian and then for $n = 3$ to conclude that $D^3$ is Noetherian and so forth. Thus $D^n$ is Noetherian.

**Proposition 9.5.4** Let $A$ be a finitely generated module for $D$ a $p.i.d$. Then $A$ is Noetherian.

**Proof:** Let $A = \text{span} (a_1, \cdots, a_n)$. Let $\tau : D^n \rightarrow A$ be given by $\tau(x) \equiv \sum_i x_i a_i$. Then this is clearly a morphism. Let $A_k$ be an ascending chain of sub modules of $A$. Then $\tau^{-1}(A_k)$ is an ascending chain of submodules of $D^n$ which contain $\ker(\tau)$. It was shown above that $D^n$ is
Noetherian. Hence there is an \( m \) such that if \( k > m, \tau^{-1}(A_k) = \tau^{-1}(A_m) \) and so \( A_k = A_m \) for all \( k > m \) showing that \( A \) is Noetherian.

Recall the elementary matrices which involved doing a row operation to the identity matrix. The elementary matrices which involve switching two rows or adding a multiple of one row to another result in elementary matrices which are invertible even if the entries come from a commutative ring with unity. See Problem 44. Similarly, these two column operations may be accomplished by multiplying on the right by an elementary matrix which involves adding a multiple of a column to another column or switching two columns. See Problem 44 on Page 111. Similarly, such elementary matrices are invertible even if the entries of the matrix come from a commutative ring.

**Theorem 9.5.5** Let \( A \) be an \( m \times n \) matrix whose entries are from a Euclidean integral domain \( D \). Thus if \( \alpha, \beta \in D \), there exists \( \kappa \) such that

\[
\alpha = \kappa \beta + \rho, \, \delta(\rho) < \delta(\beta) \quad \text{or else} \quad \rho = 0
\]

where \( \delta \) is the nonnegative integer valued function in the definition of the Euclidean ring. Then there are invertible matrices \( P, Q \) of the right size such that

\[
PAQ = B
\]

where \( B \) is a diagonal matrix. This can be done in such a way that \( B_{kk}/B_{(k+1)(k+1)} \) as long as these entries are nonzero.

**Proof:** If \( A = 0 \) there is nothing to show. Just let \( P, Q \) be appropriate identity matrices. Assume then that \( A \neq 0 \). Begin with \( P \) and \( Q \) appropriate sized identity matrices. Let \( \delta(A_{ij}) \) be the smallest of all entries of \( A \) which are not zero. Now choosing a switch of columns and rows, we can modify \( P, Q \) such that \( B_{11} = A_{ij} \). Consider \( B_{11} \), the first entry in the \( i \)th row. By the Euclidean algorithm,

\[
B_{11} = B_{11}q + r, \, \delta(r) < \delta(B_{11})
\]

or else \( r = 0 \). If it is zero, stop. If not, take \(-q\) times the first row of \( B \) and add to the \( i \)th row to place a \( r \) in the \( i1 \) position in place of \( B_{11} \). Now out of all entries of the new matrix \( B \) the \( B_{rs} \) which has \( \delta(B_{rs}) \) the smallest is in the \( i \)th row and \( \delta(B_{rs}) \leq \delta(r) \). Switch rows and columns till \( B_{is} \) is in the 11 position. Now repeat the argument just given, replacing the first entry of the \( i \)th row with a remainder \( r' \) where it is either zero or \( \delta(r') < \delta(B_{is}) \). Continuing in this way, eventually the remainder \( r \) must be zero because the process yields a strictly decreasing sequence of nonnegative integers. Now do a similar process to the other rows of the resulting matrix. When this is done, do the same thing using column operations to eventually obtain

\[
PAQ = \begin{pmatrix} B_{11} & 0^T \\ 0 & \hat{B} \end{pmatrix}
\]

If \( \hat{B} = 0 \) we are done. If not, do the same thing working with the rows of \( \begin{pmatrix} 0 & \hat{B} \end{pmatrix} \) and then the columns of

\[
\begin{pmatrix} 0^T \\ \hat{B} \end{pmatrix}
\]

to obtain

\[
PAQ = \begin{pmatrix} B_{11} & 0 \\ B_{22} & \hat{B} \end{pmatrix}
\]

where \( \hat{B} \) is now \((m-2) \times (n-2)\). Eventually, the result is a diagonal matrix.
What of the last claim that we can get \( B_{kk}/B_{(k+1)(k+1)} \)? This just involves a little more care. When you get to *, if any of the entries of \( \hat{B} \) are not a multiple of \( B_{11} \), say \( B_{ij} \), then take the column containing this entry and add to the first column. This yields

\[
\begin{pmatrix}
B_{11} & 0^T \\
* & \hat{B}
\end{pmatrix}
\]

with the offending \( B_{ij} \) the first entry of the \( i \)th row. Then do a row operation to obtain \( r \) in this position with \( 0 < \delta (r) < \delta (B_{11}) \) You can’t have \( 0 = \delta (r) \) because it is given that \( B_{ij} \) is not a multiple of \( B_{11} \). Now switch to obtain

\[
\begin{pmatrix}
r & * \\
* & \hat{B}
\end{pmatrix}
\]

Now you repeat the first part of the argument till you obtain

\[
\begin{pmatrix}
r & 0^T \\
0 & \hat{B}
\end{pmatrix}, \text{ new } \hat{B}
\]

If any entry of \( \hat{B} \) is not a multiple of \( r \), then you repeat what was just done. This process must stop because you have a sequence \( \delta (r_1) > \delta (r_2) , \cdots \) which are each positive integers with \( r_i \) in the upper left corner. Thus eventually, every entry of \( \hat{B} \) must be a multiple of \( r \) in the above. Now you do the first part of the argument to obtain

\[
\begin{pmatrix}
r_1 & 0 \\
0 & r_2 \\
0 & \hat{B}
\end{pmatrix}
\]

where \( r_1/r_2 \). The various linear combinations of entries which are multiples of \( r_1 \) will yield multiples of \( r_1 \) and so it will indeed be as shown above. Then you repeat what was just considered if \( \hat{B} \) contains an entry which is a multiple of \( r_2 \) which is a multiple of \( r_1 \). Eventually you obtain that \( B \) is a diagonal matrix and \( B_{11}/B_{22}/B_{33} \cdots \).

The above is the main result, but in fact you can generalize it to an arbitrary \textit{p.i.d.} in place of a Euclidean domain. See Jacobsen [20]. I will sketch the proof below. It is similar to the above but you have to include a slightly harder argument.

**Corollary 9.5.6** The above conclusion holds for a general \textit{p.i.d.} in place of a Euclidean domain.

**Proof:** If \( A \) is \( 1 \times 1 \), then there is nothing to show. So always it can be assumed that either \( m \) or \( n \) is larger than \( 1 \). The proof is based on the following two claims.

**Claim 1:** Suppose \( a \neq 0 \). Then there exists an invertible \( 2 \times 2 \) matrix \( R \) such that

\[
R \begin{pmatrix} a & * \\ b & *
\end{pmatrix} = \begin{pmatrix} d & * \\ 0 & *
\end{pmatrix}
\]

where \( d \) is a greatest common divisor of \( a, b \) if both are nonzero. If \( b = 0 \), you can take \( R = I \).

**Proof of the claim:** If \( a/b \), then \( b = ka \) for some \( k \) and so you can let

\[
R = \begin{pmatrix} 1 & 0 \\ -k & 1
\end{pmatrix}
\]

So assume that \( a \) does not divide \( b \) and let \( d \) be a greatest common divisor. This is unique up to associates. Then there are \( s, t \) such that

\[
d = sa + tb
\]
Now $a = rd, b = qd$. Then the above yields
\[ d = sr + tq, \quad 1 = sr + tq \]

Consider the following product.
\[
\begin{pmatrix}
  s & t \\
  -q & r
\end{pmatrix}
\begin{pmatrix}
  r & -t \\
  q & s
\end{pmatrix}
= \begin{pmatrix}
  rs + tq & -st + st \\
  -qr + rq & tq + rs
\end{pmatrix}
= \begin{pmatrix}
  1 & 0 \\
  0 & 1
\end{pmatrix}
\]

Thus both matrices are invertible. Now
\[
\begin{pmatrix}
  s & t \\
  -q & r
\end{pmatrix}
\begin{pmatrix}
  a & b \\
  * & *
\end{pmatrix}
= \begin{pmatrix}
  as + tb & * \\
  -aq + rb & *
\end{pmatrix}
= \begin{pmatrix}
  d & * \\
  -rdq + rq & *
\end{pmatrix}
= \begin{pmatrix}
  d & * \\
  0 & *
\end{pmatrix}
\]

Then analogous reasoning yields

**Claim 2:** Suppose $a \neq 0$. Then there exists an invertible $2 \times 2$ matrix $R$ such that
\[
\begin{pmatrix}
  a & b \\
  * & *
\end{pmatrix}
R = \begin{pmatrix}
  d & 0 \\
  * & *
\end{pmatrix}
\]

where $d$ is a greatest common divisor of $a$ and $b$. This is left as an exercise. See Problem 28 on Page 213.

Now the difference here is that there is no function $\delta$. Let $l(a)$ denote the number of non-invertible primes in a prime factorization of $a$. Thus, this is an integer. If $a$ is invertible, let $l(a) \equiv 0$. As before, begin with
\[
PAQ = B
\]

where $B = A$ and $P, Q$ are identity matrices which will be adjusted. Let $l(A_{ij})$ be the smallest. Adjust $P, Q$ so that $A_{ij} = B_{11}$. Now if any entries of the first column are a multiple of $B_{11}$, use an elementary matrix on the left to modify $P$ and zero these out. If $B_{i1}$ is not a multiple of $B_{11}$, move it two the 2,1 position and multiply on the left by
\[
\begin{pmatrix}
  R & 0 \\
  0 & I
\end{pmatrix}
\]
in such a way as to get a zero in this position. Keep doing this till you obtain
\[
PAQ = \begin{pmatrix}
  d_1 & a^T \\
  0 & \hat{B}
\end{pmatrix}
\]

Then use claim 2 as needed to eventually obtain
\[
PAQ = \begin{pmatrix}
  d_1 & 0^T \\
  0 & \hat{B}
\end{pmatrix}
\]

If $d_1$ fails to divide any entry of $\hat{B}$, say $B_{ij}$, add that column to the left column. Then do the above process to zero it out and replace the top left corner with the greatest common divisor of $B_{ij}$ and $d_1$. Continue doing this till you obtain a column of zeros below the top left corner. Then use claim 2 to get a row of zeros to the right of the top left corner. Thus you obtain
\[
PAQ = \begin{pmatrix}
  d_2 & 0^T \\
  0 & \hat{B}
\end{pmatrix}
\]
9.5. CYCLIC DECOMPOSITION

where $d_2/d_1$. Hence $l(d_2) < l(d_1)$. Iterate this procedure till $\hat{B}$ has no entries which are not multiples of $d_i$. This must occur because the sequence $\{l(d_i)\}$ where $d_i$ is in the upper left corner is strictly decreasing.

After this, do the same thing to $\hat{B}$ using row and column operations and multiplication by matrices of the form
\[
\begin{pmatrix}
1 & 0 \\
R & I \\
0 & 1
\end{pmatrix}
\]
Thus, eventually you obtain the same conclusion. □

Now here is the cyclic decomposition theorem.

**Theorem 9.5.7** Let $M$ be a finitely generated torsion module for a domain $D$ such that the conclusion of Theorem 9.5.5 holds. Then it is the direct sum of cyclic submodules. That is,

$$M = Dm_1 \oplus \cdots \oplus Dm_p$$

**Proof:** Let $M = Da_1 + Da_2 + \cdots + Da_n$. Since it is finitely generated, this is possible. Let $\eta : D^n \to M$

$$\eta \left( \sum_{i=1}^{n} x_i e_i \right) = \sum_{i=1}^{n} x_i a_i$$

where $e_i$ is the usual thing having a 1 in the $i^{th}$ position and zeros elsewhere. Clearly $\eta$ is a morphism. Let $K$ be the kernel of $\eta$. Let $\tilde{\eta} : D^n/K \to M$

$$\tilde{\eta} (x + K) \equiv \eta (x)$$

Thus $\tilde{\eta}$ is one to one and onto. Now it was shown above that $D^n$ is Noetherian and so $K$ is finitely generated by some $\{z_1, \ldots, z_m\}$. Let the matrix $A$ be defined by

$$z_k = \sum_{l=1}^{n} A_{kl} e_l$$

So $A$ is an $m \times n$ matrix whose entries are from $D$. Now let $Q$ be an invertible $n \times n$ matrix and define

$$e'_i = \sum_{j=1}^{n} Q_{ij} e_j, \quad \sum_{i=1}^{n} Q_{li}^{-1} e'_i = e_l$$

Also let $P$ be an invertible $m \times m$ matrix such that

$$z'_k = \sum_{j=1}^{m} P_{kj} z_j, \quad k = 1, \ldots, m \quad (***)$$

Thus

$$z'_k = \sum_{j=1}^{m} P_{kj} z_j = \sum_{j=1}^{m} P_{kj} \sum_{l=1}^{n} A_{jl} e_l$$

$$= \sum_{j=1}^{m} \sum_{l=1}^{n} P_{kj} A_{jl} \sum_{i=1}^{n} Q_{li}^{-1} e'_i$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{l=1}^{n} P_{kj} A_{jl} Q_{li}^{-1} e'_i$$

By Theorem 9.5.6, we can choose $P, Q$ such that $PAQ^{-1}$ is a diagonal. We do this now. Then the result of the above reduces to

$$z'_k = B_{kk} e'_k = d_k e'_k, \quad d_k \neq 0, k \leq m \quad (*)$$
The \( \{e_k\} \) are linearly independent in the sense that if
\[
\sum_{i=1}^{n} \alpha_i e_i = 0,
\]
then each \( \alpha_i = 0 \) because the left side is just the column vector \( \left( \alpha_1 \cdots \alpha_n \right)^T \). It is also the case that the \( \{e'_i\} \) are independent. Indeed, if you have
\[
0 = \sum_{i=1}^{n} \alpha_i e'_i
\]
then you have
\[
0 = \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{n} Q_{ij} e_j = \sum_{j=1}^{n} \left( \sum_{i=1}^{n} \alpha_i Q_{ij} \right) e_j.
\]
Hence \( \sum_{i=1}^{n} \alpha_i Q_{ij} = 0 \) for each \( j \). That is, \( \alpha^T Q = 0^T \). But \( Q^{-1} \) exists and so \( \alpha^T = 0^T \). In addition,
\[
\text{span} (e'_1 \cdots e'_n) = D^n
\]
To see this,
\[
x = \sum_{i} x_i e_i = \sum_{i} x_i \sum_{j=1}^{n} Q_{ij}^{-1} e'_j \in \text{span} (e'_1 \cdots e'_n)
\]
Now it was given that \( \text{span} (z_1, \cdots, z_m) = K \). Also \( \text{span} (z'_1, \cdots, z'_m) = K \) because by **,
\[
z'_k = \sum_{j=1}^{m} P_{kj} z_j, \quad \sum_{j=1}^{m} P_{ij}^{-1} z'_j = z_i
\]
and so anything in the span of the \( z_i \) is also in the span of the \( z'_j \). Each \( z'_k \in K \) because
\[
\eta(z'_k) = \eta \left( \sum_{j=1}^{m} P_{kj} z_j \right) = \sum_{j=1}^{m} P_{kj} \eta(z_j) = 0
\]
Now it follows that \( \{e'_{m+1} + K, \cdots, e'_n + K\} \) spans \( D^n/K \). This is because by \( \ast \), \( z'_j = d_j e'_j \) so
\[
K = \text{span} (e'_1, \cdots, e'_m)
\]
each \( e'_j \) in \( K \) for \( j \leq m \). Thus
\[
D^n/K = D (e'_{m+1} + K) + \cdots + D (e'_n + K)
\]
In fact, this is a direct sum. Suppose you have
\[
0 = \sum_{j=m+1}^{n} c_j (e'_j + K)
\]
Then you have
\[
\sum_{j=m+1}^{n} c_j e'_j = \sum_{i=1}^{m} b_i e'_i \in K
\]
for some choice of \( b_i \) and by linear independence, you have each \( c_j \) and \( b_i \) equal to zero. Thus
\[
D^n/K = D (e'_{m+1} + K) \oplus \cdots \oplus D (e'_n + K)
\]
Finally, let $\tilde{\eta}(e'_k + K) \equiv b_k$. It follows that
\[ M = Db_{m+1} + \cdots + Db_n \]

Is it a direct sum? Suppose
\[ \sum_{k=m+1}^{n} \alpha_kb_k = 0 = \sum_{k=m+1}^{n} \tilde{\eta}(\alpha_ke'_k + K) = \tilde{\eta}\left(\sum_{k=m+1}^{n} \alpha_ke'_k + K\right) \]

Hence since $\tilde{\eta}$ is one to one,
\[ \sum_{k=m+1}^{n} \alpha_ke'_k + K = 0 \]

and so each $\alpha_ke'_k + K = 0$ in $D^n/K$ and so $\alpha_ke'_k \in K$ for each $k$ and so $\tilde{\eta}(\alpha_ke'_k + K) \equiv \eta(\alpha_ke'_k) = 0$ this for each $k$. Thus $\alpha_kb_k = 0$ and so this is indeed a direct sum. Hence $M = Db_{m+1} + \cdots + Db_n$.

Now recall the following theorem, Theorem 9.5.8.

**Theorem 9.5.8** Let $M$ be a non zero torsion module for a p.i.d. $D$ and suppose that
\[ M = Dz_1 + \cdots + Dz_a \]

so that it is finitely generated. Then there exist primes $\{p_i\}_{i=1}^n$ distinct in the sense that none is an invertible multiple of another such that
\[ M = M_{p_1} \oplus \cdots \oplus M_{p_n}, \text{ no } M_{p_i} \text{ equal to } 0 \]

For $(\beta) = \ann(M), \beta = \prod_{i=1}^{n} p_i^{k_i},$ a product of non-invertible primes, it follows
\[ \tilde{M}_{p_i} \equiv \left\{ m \in M : p_i^{k_i}m = 0 \right\} = \left\{ m \in M : p_i^{k_i}m = 0 \right\} \text{ for some } k \}

which is not dependent on the spanning set.

Now, the sub modules $M_p$ in the above theorem will be decomposed into direct sums of cyclic sub modules. Pick one of the $p_j$ and consider $M_{p_j}$ which, by the above, is of the form
\[ \left\{ m \in M : p_j^{k_j}m = 0 \right\} \text{ for a fixed } k_j. \]

By Theorem 9.5.8, $M_{p_j}$ can be decomposed into a direct sum.
\[ M_{p_j} = Da_{j_1} + \cdots + Da_{j_{m_j}} \]

Now $\ann(a_{j_1}) = (\nu)$ for some $\nu$ since this is a torsion module. Hence $\nu/p_j^{k_j}$ and so $\nu = p_j^{l_j}$ for some $l_j \leq k_j$. We can arrange the order of the direct summands such that $l_1 \leq l_2 \leq \cdots \leq l_{m_j}$. Thus the following is obtained.

**Theorem 9.5.9** Let $M$ be a non zero torsion module for a p.i.d. $D$ and suppose that
\[ M = Dz_1 + \cdots + Dz_p \]

so that it is finitely generated. Then there exist primes $\{p_i\}_{i=1}^n$ distinct in the sense that none is an invertible multiple of another such that
\[ M = M_{p_1} \oplus \cdots \oplus M_{p_n}, \text{ no } M_{p_i} \text{ equal to } 0 \]

For $(\beta) = \ann(M), \beta = \prod_{i=1}^{n} p_i^{k_i},$ a product of non-invertible primes, it follows
\[ \tilde{M}_{p_i} \equiv \left\{ m \in M : p_i^{k_i}m = 0 \right\} = \left\{ m \in M : p_i^{k_i}m = 0 \right\} \text{ for some } k \}

which is not dependent on the spanning set. Also,
\[ M_{p_j} = Da_{j_1} + \cdots + Da_{j_{m_j}}, m_j \leq k_j \]

where $\ann(a_{j_r}) = p_j^{l_r}$, such that $l_r \leq k_j$ and we can have the order arranged such that $l_1 \leq l_2 \leq \cdots \leq l_{m_j} \leq k_j$. 

Note that the proof shows that each \( M_{p_i} \) is finitely generated. Also, \( \text{ann}(M) \) implies that \( l_{m_j} = k_j \) since otherwise, we could take \( k_j \) smaller and there would be no change so \((\beta)\) would not really equal \( \text{ann}(M) \).

There is a question about whether the \( l_i \) arranged in increasing (decreasing) order are unique. This is because the cyclic decomposition in the above theorem is not unique.

### 9.6 Uniqueness

The following discussion follows \[44\].

Suppose you have a finitely generated torsion module \( M, \text{ann}(M) = (p^\theta) \) for \( p \) a prime in the \( p.i.d.D \). Also let

\[ M = Dv_1 \oplus \cdots \oplus Dv_s, \text{ no } v_j = 0 \]

where these are numbered such that \( \text{ann}(Dv_1) \supseteq \text{ann}(Dv_2) \supseteq \cdots \supseteq \text{ann}(Dv_s) \). Thus \( \text{ann}(v_j) = (p^{l_j}) \) where \( l_j \leq q \). If this is numbered so that the annihilators are ordered as just described, then \( l_1 \leq l_2 \leq \cdots \leq l_s \leq q \). Since \( \text{ann}(M) = p^q \), it must be the case that \( l_s = q \) since if not, \( \text{ann}(M) \neq (p^q) \). It would equal to \( (p^{l_j}) \) instead. By Proposition \[45\], \( D/(p) = \hat{D} \) is a field.

In addition, we have \( M \supseteq pM \supseteq p^2M \supseteq \cdots \). These are each submodules of \( M \).

**Claim:** We can consider \( p^kM/p^{k+1}M \) as a vector space over \( \hat{D} \) as follows.

\[ (\alpha + (p))(m + p^{k+1}M) = \alpha m + p^{k+1}M \]

**Proof of claim:** This operation needs to be well defined and then it will be obvious that indeed this is a vector space. Suppose then that \( \alpha + (p) = \hat{\alpha} + (p) \) and \( m + p^{k+1}M = \hat{m} + p^{k+1}M \). Is it the case that \( \alpha m - \hat{\alpha} \hat{m} \in p^{k+1}M \)? The difference equals

\[ \alpha (m - \hat{m}) + (\alpha - \hat{\alpha}) \hat{m} \]

The first term is in \( p^{k+1}M \) because it is given that \( m - \hat{m} \) is in \( p^{k+1}M \). Consider the second term. It is given that \( \hat{m} = p^k\hat{m}_1 \) because it is in \( p^kM \). Also, it is given that \( \alpha - \hat{\alpha} \) is a multiple of \( p \) and so indeed the second term is also in \( p^{k+1}M \). This proves the claim.

If \( k \geq l_s = q \), then \( p^kM = 0 \) and so for such \( k, p^kM/p^{k+1}M = 0 \). The other case is that \( k < l_s \). Say \( k \in [l_j, l_{j+1}) \). Then

\[ p^kM = Dp^k v_{j+1} + Dp^k v_{j+2} + \cdots + Dp^k v_s. \]

This is because if \( k \geq l_j \), then for \( i \leq j, p^k v_i = 0 \) and so all that survives is what is in the above sum.

Is

\[ \{p^k v_{j+1} + p^{k+1} M, \ldots, p^k v_s + p^{k+1} M \} \]

linearly independent over \( \hat{D} \)? Suppose

\[ \sum_{r=j+1}^s (\alpha_r + (p))(p^k v_r + p^{k+1}M) = 0 \]

Then

\[ \sum_{r=j+1}^s \alpha_r p^k v_r + p^{k+1}M = 0 \]

This requires

\[ \sum_{r=j+1}^s \alpha_r p^k v_r = p^{k+1}M \]

for some \( m \in M \). However, recall that \( M \) is a direct sum and so

\[ \sum_{r=j+1}^s \alpha_r p^k v_r = \sum_{r=j+1}^s p^{k+1} \beta_r v_r \]
Thus, since $M$ is a direct sum again, for each $r$,

$$p^k \alpha_r = p^{k+1} \beta_r$$

and so $\alpha_r$ is a multiple of $p$. Hence $\alpha_r + (p) = 0$ for each $r$ and this is indeed a basis for $p^k M/p^{k+1}M$ over $D$. It follows that the dimension of $p^k M/p^{k+1}M$ over $D$ is $s - j$ where $k \in [l_j, l_{j+1})$, the number of $v_{j+1}$ for $l_{j+1} > k$.

Now suppose that

$$M = Dw_1 \oplus \cdots \oplus Dw_l$$

such that $\text{ann} (Dw_1) \supseteq \text{ann} (Dw_2) \supseteq \cdots \supseteq \text{ann} (Dw_l)$ and $\text{ann} (w_j) = (p^{m_j})$. The question is whether $s = t$ and $m_j = l_j$. It was just shown that for $k$ a positive integer, the dimension of $p^k M/p^{k+1}M$ is the number of $v_j$ for $l_j > k$. Similarly it is the number of $w_j$ for $m_j > k$ and this must be the same number. In other words, for each $k$ there are the same number of $m_j$ larger than $k$ as there are $l_j$ larger than $k$. Thus the $l_j$ coincide with the $m_j$ and there are the same number of them. Hence $s = t$ and $m_j = l_j$.

This last assertion deserves a little more explanation. The smallest $l_j$ and $k_j$ can be is 1. Otherwise, you would have $\text{ann} (w_1)$ or $\text{ann} (v_1) = (p^a)$, but 1. Thus for every $\alpha \in D$, $\alpha w_1 = 0$. In particular, this would hold if $\alpha = 1$ and so $w_l = 0$ contrary to assumption that none of the $w_i, v_i$ equal to 0. (Obviously you can’t get uniqueness if you allow some $v_j$ to equal 0 because you can string together as many $D0$ as you like.) Therefore, you could consider $M/pM$ such that $k = 0$ and there are the same number of $v_i$ and $w_j$ since each $l_i, m_j$ is larger than 0. Thus $s = t$. Then a contradiction is also obtained if some $m_i \neq l_i$. You just consider the first such $i$ and let $k$ be the smaller of the two.

**Theorem 9.6.1** Suppose $M$ is a finitely generated torsion module for a p.i.d. and $\text{ann} (M) = (p^a)$ where $p$ is a prime. Then

$$M = Dv_1 \oplus \cdots \oplus Dv_s, \text{ no } v_j = 0$$

Letting $\text{ann} (v_j) = (\nu_j)$, it follows that $\nu = p^b$ for some $l_j \leq q$. If the direct summands are listed in the order that the $l_i$ are increasing (or decreasing), then $s$ is independent of the choice of the $v_j$ and any other such cyclic direct sum for $M$ will have the same sequence of $l_j$.

These considerations about uniqueness yield the following more refined version of Theorem 9.6.2.

**Theorem 9.6.2** Let $M$ be a non zero torsion module for a p.i.d. $D$ and suppose that

$$M = Dz_1 + \cdots + Dz_p$$

so that it is finitely generated. Then there exist primes \{p_i\}_{i=1}^n distinct in the sense that none is an invertible multiple of another such that

$$M = M_{p_1} \oplus \cdots \oplus M_{p_n}, \text{ no } M_{p_i} \text{ equal to } 0$$

For $(\beta) = \text{ann} (M), \beta = \prod_{i=1}^n p_i^{k_i}$, a product of non-invertible primes, it follows

$$M_{p_i} \equiv \left\{ m \in M : p_i^{k_i} m = 0 \right\} = \left\{ m \in M : p_i^{k_i} m = 0 \text{ for some } k \right\} = M_{p_i}$$

which is not dependent on the spanning set. Also,

$$M_{p_i} = Da_{j_1} \oplus \cdots \oplus Da_{j_{m_j}}, m_j \leq k_j$$

where $\text{ann} (a_{j_r}) = p^{l_r}$, such that $l_r \leq k_j$ and we can have the order arranged such that $l_1 \leq l_2 \leq \cdots \leq l_{m_j} = k_j$. The numbers

$$l_1 \leq l_2 \leq \cdots \leq l_{m_j} = k_j$$

are the same for each cyclic decomposition of $M_{p_i}$. That is, they do not depend on the particular decomposition chosen.
CHAPTER 9. MODULES AND RINGS

Of course we would also have uniqueness if we adopted the convention that the $l_i$ should be arranged in decreasing order.

All of the above is included in [23] and 1 where, in addition to the above, there are some different presentations along with much more on modules and rings.

9.7 Canonical Forms

Now let $D = F[x]$ and let $L \in \mathcal{L}(V,V)$ where $V$ is a finite dimensional vector space over $F$. Define the multiplication of something in $D$ with something in $V$ as follows.

$$g(x)v \equiv g(L)v, \; L^0 \equiv I$$

As mentioned above, $V$ is a finitely generated module and $F[x]$ is a $p.i.d$. The non-invertible primes are polynomials irreducible over $F$. Letting $\{z_1, \cdots, z_l\}$ be a basis for $V$, it follows that

$$V = Dz_1 + \cdots + Dz_l$$

and so there exist irreducible polynomials over $F$, $\{\phi_i(x)\}_{i=1}^n$ and corresponding positive integers $k_i$ such that

$$V = \ker(\phi_1(L)^{k_1}) \oplus \cdots \oplus \ker(\phi_n(L)^{k_n})$$

This is just Theorem 9.7.1 obtained as a special case. Recall that the entire theory of canonical forms is based on this result. This follows because

$$M_{\phi_i} = \{v \in V : \phi_i(x)^{k_i}v = \phi_i(L)^{k_i}v = 0\}.$$ 

Now continue letting $D = F[x], L \in \mathcal{L}(V,V)$ for $V$ a finite dimensional vector space and $g(x)v \equiv g(L)v$ as above. Consider $M_{\phi_i} = \ker(\phi_i(L)^{k_i})$ which is a sub module of $V$. Then by Theorem 9.7.1, this is the direct sum of cyclic sub modules.

$$M_{\phi_i} = Dv_1 \oplus \cdots \oplus Dv_s \quad (*)$$

where $s = \text{rank}(M_{\phi_i})$.

At this point, note that $\text{ann}(M_{\phi_i}) = (\phi_i(x)^{k_i})$ and so $\text{ann}(Dv_j) = (\phi_i(x)^{l_j})$ where $l_j \leq k_i$. If $d_i$ is the degree of $\phi_i(x)$, this implies that for $v_j$ being one of the $v$ in $*$,

$$1, L v_j, L^2 v_j, \cdots, L^{l_j d_i - 1} v_j$$

must be a linearly independent set since otherwise, $\text{ann}(Dv_j) \neq (\phi_i(x)^{l_j})$ because a polynomial of smaller degree than the degree of $\phi_i(x)$ will annihilate $v_j$. Also,

$$Dv_j \subseteq \text{span}(v_j, L v_j, L^2 v_j, \cdots, L^{l_j d_i - 1} v_j)$$

by division. Vectors in $Dv_j$ are of the form $g(x)v_j = (\phi_i(x)^{l_j} k(x) + \rho(x)) v_j = \rho(x) v_j$ where the degree of $\rho(x)$ is less than $l_j d_i$. Thus a basis for $Dv_j$ is \{v_j, L v_j, L^2 v_j, \cdots, L^{l_j d_i - 1} v_j\} and the span of these vectors equals $Dv_j$. Now you see why the term "cyclic" is appropriate for the submodule $Dv$. This shows the following theorem.

**Theorem 9.7.1** Let $V$ be a finite dimensional vector space over a field of scalars $F$. Also suppose the minimum polynomial is $\prod_{i=1}^n (\phi_i(x))^{k_i}$, where $k_i$ is a positive integer and the degree of $\phi_i(x)$ is $d_i$. Then

$$V = \ker(\phi_1(L)^{k_1}) \oplus \cdots \oplus \ker(\phi_n(L)^{k_n})$$

$$\equiv M_{\phi_1(x)} \oplus \cdots \oplus M_{\phi_n(x)}$$
Furthermore, for each \( i \), in \( \ker (\phi_i (L)^{k_i}) \), there are vectors \( v_1, \ldots, v_s \) and positive integers \( l_1, \ldots, l_s \), each no larger than \( k_i \) such that a basis for \( \ker (\phi_i (L)^{k_i}) \) is given by

\[
\{ \beta_{v_1}^{l_1}, \ldots, \beta_{v_s}^{l_s}\}
\]

where the symbol \( \beta_{v_j}^{l_j} \) signifies the ordered basis

\[
(v_j, L v_j, L^2 v_j, \ldots, L^{l_j-2} v_j, L^{l_j-1} v_j)
\]

Its length is the degree of \( \phi_j (x)^{k_j} \) and is therefore, determined completely by the \( l_j \). Thus the lengths of the \( \beta_{v_j}^{l_j} \) are uniquely determined if they are listed in order of increasing or decreasing length.

The last claim of this theorem will mean that the various canonical forms are uniquely determined.

It is clear that the span of \( \beta_{v_j}^{l_j} \) is invariant with respect to \( L \) because, as discussed above, this span is \( D v_j \) where \( D = \mathbb{F} [x] \). Also recall that \( \text{ann} (D v_j) = (\phi_i (x)^{l_j}) \) where \( l_j \leq k_i \). Let

\[
\phi_i (x)^{l_j} = x^{l_j} + a_{n-1} x^{l_j-1} + \cdots + a_1 x + a_0
\]

Recall that the minimum polynomial has leading coefficient equal to 1. Of course this makes no difference in the above presentation because \( a_n \) is invertible and so the ideals and above direct sum are the same regardless of whether this leading coefficient equals 1, but it is convenient to let this happen since otherwise, the blocks for the rational canonical form will not be standard. Then what is the matrix of \( L \) restricted to \( D v_j \)?

\[
\left( \begin{array}{ccc}
Lv_j & \cdots & L^{l_j-1} v_j \\
\cdots & \cdots & \cdots \\
L^{l_j-1} v_j & & L^{l_j-1} v_j
\end{array} \right) = \left( \begin{array}{ccc}
v_j & \cdots & L^{l_j-2} v_j \\
\cdots & \cdots & \cdots \\
L^{l_j-1} v_j & & L^{l_j-1} v_j
\end{array} \right) M
\]

where \( M \) is the desired matrix. Now \( \text{ann} (D v_j) = (\phi_i (x)^{l_j}) \) and so

\[
L^{l_j} v_j = (-1) (a_{n-1} L^{l_j-1} v_j + \cdots + a_1 L v_j + a_0 v_j)
\]

Thus the matrix \( M \) must be of the form

\[
\begin{pmatrix}
0 & -a_0 \\
1 & -a_1 \\
\vdots & \vdots \\
0 & 1 -a_{n-1}
\end{pmatrix}
\]

(9.2)

It follows that the matrix of \( L \) with respect to the basis obtained as above will be a block diagonal with blocks like the above. This is the rational canonical form.

Of course, those blocks corresponding to \( \ker (\phi_i (L)^{k_i}) \) can be arranged in any order by just listing the \( \beta_{v_1}^{l_1}, \ldots, \beta_{v_s}^{l_s} \) in various orders. If we want the blocks to be larger in the top left and get smaller towards the lower right, we just re-number it to have \( l_i \) be a decreasing sequence. Note that this is the same as saying that \( \text{ann} (D v_1) \subseteq \text{ann} (D v_2) \subseteq \cdots \subseteq \text{ann} (D v_s) \). If we want the blocks to be increasing in size from the upper left corner to the lower right, this corresponds to re-numbering such that \( \text{ann} (D v_1) \supseteq \text{ann} (D v_2) \supseteq \cdots \supseteq \text{ann} (D v_1) \). This second one involves letting \( l_1 \leq l_2 \leq \cdots \leq l_s \).

What about uniqueness of the rational canonical form given an order of the spaces \( \ker (\phi_i (L)^{k_i}) \) and under the convention that the blocks associated with \( \ker (\phi_i (L)^{k_i}) \) should be increasing or degreasing in size from upper left toward lower right? In other words, suppose you have

\[
\ker (\phi_i (L)^{k_i}) = D v_1 \oplus \cdots \oplus D v_s = D w_1 \oplus \cdots \oplus D w_t
\]
such that
\[ \text{ann}(Dv_1) \subseteq \text{ann}(Dv_2) \subseteq \cdots \subseteq \text{ann}(Dv_s) \]
and
\[ \text{ann}(Dw_1) \subseteq \text{ann}(Dw_2) \subseteq \cdots \subseteq \text{ann}(Dw_t) \]
will it happen that \( s = t \) and the that the blocks associated with corresponding \( Dv_i \) and \( Dw_i \) are the same size? In other words, if \( \text{ann}(Dv_j) = (\phi_j(x)^{m_j}) \) and \( \text{ann}(Dw_j) = (\phi_j(x)^{m_{j'}}) \) is \( m_j = l_j \).

If this is so, then this proves uniqueness of the rational canonical form up to order of the blocks. However, this was proved above in the discussion on uniqueness, Theorem 9.6.1.

In the case that the minimum polynomial splits the following is also obtained.

**Corollary 9.7.2** Let \( V \) be a finite dimensional vector space over a field of scalars \( F \). Also let the minimal polynomial be \( \prod_{i=1}^n (x - \mu_i)^{k_i} \) where \( k_i \) is a positive integer. Then
\[ V = \ker((L - \mu_1 I)^{k_1}) \oplus \cdots \oplus \ker((L - \mu_n I)^{k_n}) \]
Furthermore, for each \( i \), in \( \ker((L - \mu_i I)^{k_i}) \), there are vectors \( v_1, \ldots, v_{s_i} \) and positive integers \( l_1, \ldots, l_{s_i} \), each no larger than \( k_i \) such that a basis for \( \ker((L - \mu_i I)^{k_i}) \) is given by
\[ \{\beta_{v_1}^{l_1-1}, \ldots, \beta_{v_{s_i}}^{l_{s_i}-1}\} \]
where the symbol \( \beta_{v}^{j-1} \) signifies the ordered basis
\[ ((L - \mu_i I)^{l_j-1} v_j, (L - \mu_i I)^{l_j-2} v_j, \ldots, (L - \mu_i I)^2 v_j, (L - \mu_i I) v_j, v_j) \]
(Note how this is the reverse order to the above. This is to follow the usual convention in the Jordan form in which the string of ones is on the super diagonal.)

**Proof:** The proof is essentially the same.
\[ \ker((L - \mu_1 I)^{k_1}) = Dv_1 \oplus \cdots \oplus Dv_{s_1} \]
\[ \text{ann}(Dv_j) \] for \( v_j \in \ker((L - \mu_i I)^{k_i}) \) is a principal ideal \( (\nu(x)) \) where \( \nu(x)/(x - \mu_i)^{k_i} \) and so it is of the form \((x - \mu_i)^{l_j}\) \(0 \leq l_j \leq k_i\). Then as before,
\[ v_j, (L - \mu_i I) v_j, (L - \mu_i I)^2 v_j, \ldots, (L - \mu_i I)^{l_j-1} v_j \]

(*) must be independent because if not, there is a polynomial \( g(x) \) of degree less than \( l_j \) such that \( g(x)(Dv_j) = 0 \) and so \((x - \mu_i)^{l_j}\) cannot really be \( \text{ann}(Dv_j) \). It is also true that the above list of vectors must span \( Dv_j \) because if \( f(x) \in D \), then \( f(x) = (x - \mu_i)^{l_j} m(x) + r(x) \) where \( r(x) \) has degree less than \( l_j \). Thus \( f(x) v_j = r(x) v_j \) and clearly \( r(x) \) can be written in the form \( r(x - \mu_i) \) with the same degree. (Just take Taylor expansion of \( r \) formally.) Thus * is indeed a basis of \( Dv_j \) for \( v_j \in \ker((L - \mu_i I)^{k_i}) \). \( \blacksquare \)

This gives the Jordan form right away. In this case, we know that \((L - \mu_i I)^{l_j} v_j = 0\) and so the matrix of the transformation \( L - \mu_i I \) with respect to this basis on \( Dv_j \) obtained in the usual way.
\[
\begin{pmatrix}
0 & (L - \mu_i I)^{l_j-1} v_j & \cdots & (L - \mu_i I) v_j \\
(L - \mu_i I)^{l_j-1} v_j & (L - \mu_i I)^{l_j-2} v_j & \cdots & v_j \\
\end{pmatrix}
\]
9.8 Exercises

1. Explain why any finite Abelian group is a module over the integers. Explain why every finite Abelian group is the direct sum of cyclic subgroups.

2. Let \( R \) be a commutative ring which has 1. Show it is a field if and only if the only ideals are \((0)\) and \((1)\).

3. Let \( R \) be a commutative ring and let \( I \) be an ideal of \( R \). Let

\[ N(I) = \{ x \in R : x^n \in I \text{ for some } n \in \mathbb{N} \}. \]

Show that \( N(I) \) is an ideal containing \( I \). Also show that \( N(N(I)) = N(I) \). This is called the radical of \( I \). Hint: If \( x^n \in I \) for some \( n \), what of \( xy \)? Is it true that \((xy)^n \in I\)? For the second part, it is clear that \( N(N(I)) \supseteq N(I) \). If \( x \in N(N(I)) \), then \( x^n \in N(I) \) for some \( n \). Is it true that \((x^n)^k = x^{kn} \in I \) for some \( k \)?

4. Let \( \mathbb{F} \) be a field and \( p(x) \in \mathbb{F}[x] \). Consider \( R = \mathbb{F}[x]/(p(x)) \). Let

\[ N = \{ q \in R : q^n = 0 \text{ for some } n \in \mathbb{N} \} \]

Show that \( N \) is an ideal and that it equals \((0)\) if and only if \( p(x) \) is not divisible by the square of any polynomial. Hint: Say \( N = (0) \). Then by assumption, if \( q(x)^n \in (p(x)) \), we must have \( q(x) \) a multiple of \( p(x) \). Say \( p(x) = \prod_{i=1}^{m} p_i(x)^{k_i} \). Argue that a contradiction results if any \( k_i > 1 \) by replacing \( k_i \) in the product with \( l_i, 1 \leq l_i < k_i \). This is \( q(x) \). Explain why \( q(x)^n \in (p(x)) \) but \( q(x) \notin (p(x)) \). Therefore, \( p(x) = \prod_{i=1}^{m} p_i(x) \), the \( p_i(x) \) distinct noninvertible primes. Explain why this precludes having \( p(x) \) divisible by \( q(x)^2 \). Conversely, if \( p(x) \) is divisible by \( q(x)^2 \) for some polynomial, then \( p(x) = \prod_{i=1}^{m} p_i(x)^{k_i} \) where some \( k_i > 1 \) and so \( N \neq (0) \). Explain.

5. Recall that if \( p(x) \) is irreducible in \( \mathbb{F}[x] \), then \( \mathbb{F}[x]/(p(x)) \) is a field. Show that if \( p(x) \) is not irreducible, then this quotient ring is not even an integral domain.

6. Find a polynomial \( p(x) \) of degree 3 which is irreducible over \( \mathbb{Z}_3 \). Thus is a prime in \( \mathbb{Z}_3[x] \). Now consider \( \mathbb{Z}_3[x]/(p(x)) \). Recall that this is a field. How many elements are in this field? Hint: Show that all elements of this field are of the form \( a_0 + a_1 x + a_2 x^2 + (p(x)) \) where there are three choices for each \( a_i \).

7. Let \( \mathbb{F} \) be a field and consider \( \mathbb{F}[x]/(p(x)) \) where \( p(x) \) is prime in \( \mathbb{F}[x] \) (irreducible). Then this was shown to be a field. Let \( \alpha \) be a root of \( p(x) \) in a field \( \mathbb{G} \) which contains \( \mathbb{F} \). Consider \( \mathbb{F}[\alpha] \) which means all polynomials in \( \alpha \) having coefficients in \( \mathbb{F} \). Show that this is isomorphic to \( \mathbb{F}[x]/(p(x)) \) and is itself a field. Hint: Let \( \theta(k(x) + (p(x))) \equiv k(\alpha) \). Show that since \( p(x) \) is

a Jordan block for the nilpotent matrix \((L - \mu_i I)\). Thus, with respect to this basis, the block associated with \( L \) and \( \beta_{ij}^{l_j} \) is just

\[
\begin{pmatrix}
\mu_i & 1 & 0 \\
\mu_i & \ddots & \ddots \\
0 & \ddots & \ddots & 1 \\
& \ddots & \ddots & \ddots & 1 \\
\end{pmatrix}
\]

This has proved the existence of the Jordan form. You have a string of blocks like the above for \( \ker(L - \mu_i I)^{k_i} \). Of course, these can be arranged so that the size of the blocks is decreasing from upper left to lower right. As with the rational canonical form, once it is decided to have the blocks be decreasing (increasing) in size from upper left to lower right, the Jordan form is unique.
irreducible, it has the smallest possible degree out of all polynomials \( q(x) \) for which \( q(\alpha) = 0 \). You might use this to show that \( \theta \) is one to one. Show that \( \theta \) is actually an isomorphism. Thus \( \mathbb{F}[\alpha] \) must be a field.

8. Letting \( \mathbb{Z} \) be the integers, show that \( \mathbb{Z}[x] \) is an integral domain. Is this a principal ideal domain? Show that in \( \mathbb{Z}[x] \), if you have \( f(x) \), \( g(x) \) given, then there exists an integer \( b \) such that

\[
b f(x) = g(x) Q(x) + R(x)
\]

where the degree of \( R(x) \) is less than the degree of \( g(x) \). Note how in a field, you don’t need to multiply by some integer \( b \). **Hint:** Concerning the question about whether this is a p.i.d., suppose you have an ideal in \( \mathbb{Z}[x] \) called \( I \). Say \( l(x) \in I \) and \( l(x) \) has smallest degree out of all polynomials in \( I \). Let \( k(x) \in I \). Then in \( \mathbb{Q}[x] \), you have \( r(x), q(x) \in \mathbb{Q}[x] \) such that

\[
k(x) = l(x) q(x) + r(x),
\]

\[
r(x) = 0 \text{ or degree of } r(x) < \text{ degree of } l(x)
\]

Now multiply by denominators to get an equation in which everything is in \( \mathbb{Z}[x] \) but all the degrees are the same. Let the result on the left be \( \hat{k}(x) \). Then it is in \( I \). Obtain a contradiction if \( r(x) \neq 0 \).

9. Now consider the polynomials \( x^3 + x \) and \( 2x^2 + 1 \). Show that you cannot write

\[
x^3 + x = q(x) (2x^2 + 1) + r(x)
\]

where the degree of \( r(x) \) is less than the degree of \( 2x^2 + 1 \) and both \( r(x), q(x) \) are in \( \mathbb{Z}[x] \). Thus, even though \( \mathbb{Z}[x] \) is a p.i.d., the degree will not serve to make \( \mathbb{Z}[x] \) into a Euclidean domain.

10. The symbol \( \mathbb{F}[x_1, \ldots, x_n] \) denotes the polynomials in the indeterminates \( x_1, \ldots, x_n \) which have coefficients in the field \( \mathbb{F} \). Thus a typical element of \( \mathbb{F}[x_1, \ldots, x_n] \) is of the form

\[
\sum_{i_1, \ldots, i_n} a_{i_1, \ldots, i_n} x_1^{i_1} \cdots x_n^{i_n}
\]

where this sum is taken over all lists of nonnegative integers \( i_1, \ldots, i_n \) and only finitely many \( a_{i_1, \ldots, i_n} \in \mathbb{F} \) are nonzero. Explain why this is a commutative ring which has 1. Also explain why this cannot be a principal ideal ring in the case that \( n > 1 \). **Hint:** Consider \( \mathbb{F}[x, y] \) and show that \( (x, y) \), denoting all polynomials of the form \( ax + by \) where \( a, b \in \mathbb{F} \) cannot be obtained as a principal ideal \( (p(x, y)) \).

11. Suppose you have a commutative ring \( R \) and an ideal \( I \). Suppose you have a morphism \( h : R \to \tilde{R} \) where \( \tilde{R} \) is another ring and that \( I \subseteq \ker(h) \). Also let

\[
f : R \to R/I, \text{ be defined by } f(r) \equiv r + I
\]

Here \( R/I \) is as described earlier in the chapter. The entries are of the form \( r + I \) where \( r \in R \).

\[
\begin{align*}
    r + I + \hat{r} + I & \equiv r + \hat{r} + I \\
n(r + I) (\hat{r} + I) & \equiv r\hat{r} + I
\end{align*}
\]

show this is a ring and that the operations are well defined. Show that \( f \) is a morphism and is also onto. Then show that there exists \( \theta : R/I \to \tilde{R} \) such that \( h = \theta \circ f \)

\[
\begin{array}{ccc}
R & \xrightarrow{f} & R/I \\
\downarrow & \searrow \theta & \downarrow \theta \\
\tilde{R}
\end{array}
\]

**Hint:** It is clear that \( f \) is well defined and onto and is a morphism. Define \( \theta(r + I) \equiv h(r) \).

Now show it is well defined.
12. The Gaussian integers \( \mathbb{Z}[i] \) are complex numbers of the form \( m + in \) where \( m, n \) are integers. Show that this is also the same as polynomials with integer coefficients in powers of \( i \) which explains the notation. Show this is an integral domain. Reviewing the Definition \[ \] show that the Gaussian integers are also a Euclidean domain if \( \delta(a) \equiv a\bar{a} \).\textbf{ Hint: }For the last part, if you have \( a, b \in \mathbb{Z}[i] \), then as in \[ \]
\[
\frac{a}{b} = \mu + i\lambda
\]
where \( \mu, \nu \in \mathbb{Q} \) the rational numbers. There are integers \( u, l \) such that \( |\mu - u| \leq 1/2, |\lambda - l| \leq 1/2 \). Then letting \( \varepsilon \equiv \mu - u, \eta \equiv \lambda - l \), it follows that \( a = b(u + \varepsilon) + i(l + \eta) \)
Thus \( a = b(u + il) + b(\varepsilon + i\eta) \)
It follows that \( r \in \mathbb{Z}[i] \). Why? Now consider \( \delta(r) \equiv b(\varepsilon + i\eta)b(\varepsilon + i\eta) = b(\varepsilon + i\eta)b(\varepsilon - i\eta) \)
Verify that \( \delta(r) < \delta(b) \). Explain why \( r \) is maybe not unique.

13. This, and the next several problems give a more abstract treatment of the cyclic decomposition theorem found in Birkhoff and Mcclain.\[ \]
Suppose \( B \) is a submodule of \( A \) a torsion module over a \( p.i.d.D \). Show that \( A/B \) is also a module over \( D \), and that \( \text{ann} (A/B) \) equals \( \langle \alpha \rangle \) for some \( \alpha \in D \). Now suppose that \( A/B \) is cyclic \( A/B = Da_0 + B \)
Say \( \langle \beta \rangle = \text{ann} (a_0) \). Now define \( I \equiv \{ \lambda : \lambda a_0 \in B \} \). Explain why this is an ideal which is at least as large as \( \langle \beta \rangle \). Let it equal \( \langle \alpha \rangle \). Explain why \( \alpha/\beta \). In particular, \( \alpha \neq 0 \). Then let \( \theta : D \to B \oplus D \) be given by \( \theta(\lambda) \equiv (\alpha \lambda a_0, -\lambda \alpha) \).
Here we write \( B \oplus D \) to indicate \( B \times D \) but with a summation and multiplication by a scalar defined as follows: \( (b, \alpha) + (\bar{b}, \bar{\alpha}) \equiv (\bar{b} + b, \alpha + \bar{\alpha}) \), \( \beta(a, \lambda) \equiv (\beta a, \beta \lambda) \). Show that this mapping is one to one and a morphism meaning that it preserves the operation of addition, \( \theta(\delta + \lambda) = \theta \delta + \theta \lambda \). Why is \( \alpha \lambda a_0 \in B \)?

14. \[\]
In the above problem, define \( \eta : B \oplus D \to A \) as \( \eta(b, \alpha) \equiv b + \alpha a_0 \). Show that \( \eta \) maps onto \( A \), that \( \eta \circ \theta = 0 \), and that \( \theta(D) = \ker(\eta) \). People write the result of these two problems as \( 0 \to D \xrightarrow{\theta} B \oplus D \xrightarrow{\eta} A \to 0 \)
and it is called a “short exact sequence”. It is exact because \( \theta(D) = \ker(\eta) \).

15. \[\]
Let \( C \) be a cyclic module with \( \text{ann} (C) = \langle \mu \rangle \) and \( \mu A = 0 \). Also, as above, let \( A \) be a module with \( B \) a submodule such that \( A/B \) is cyclic, \( A/B \equiv D(a_0 + B), \text{ann} (A/B) = \langle \nu \rangle, C = Da_0 \). From the above problem, \( \nu \neq 0 \). Suppose there is a morphism \( \tau : B \to C \). Then it is desired to extend this to \( \sigma : A \to C \) as illustrated in the following picture where in the picture, \( i \) is the inclusion map.

\[
\begin{array}{ccc}
B & \xrightarrow{i} & A \\
\uparrow & & \uparrow \\
C & \xleftarrow{\tau} & \sigma
\end{array}
\]
You need to define \( \sigma \). First explain why \( \mu = \nu \lambda \) for some \( \lambda \). Next explain why \( \nu a_0 \in B \) and then why this implies \( \lambda \tau (\nu a_0) = 0 \). Let \( \tau (\nu a_0) = \beta c_0 \). The \( \beta \) exists because of the assumption
that \( \tau \) maps to \( C = Dc_0 \). Then since \( \lambda \tau (\nu a_0) = 0 \), it follows that \( \lambda \beta c_0 = 0 \). Explain why \( \lambda \beta = \mu \delta = \nu \lambda \delta \). Then explain why \( \beta = \nu \delta \).

Thus

\[
\begin{align*}
\tau (\nu a_0) &= \nu \delta c_0 = \nu c' \\
\tau (\kappa \nu a_0) &= \kappa \nu c'
\end{align*}
\]

From the above problem,

\[
0 \to D \xrightarrow{\theta} B \oplus D \xrightarrow{\eta} A \to 0
\]

where \( \theta (\kappa) \equiv (\kappa \nu a_0, -\kappa \nu) \), \( \eta (b, \kappa) \equiv b + \kappa a_0 \) and the above is a short exact sequence. So now extend \( \tau \) to \( \tau' : B \oplus D \to C \) as follows.

\[
\tau' (b, \kappa) \equiv \tau (b) + \kappa c'
\]

The significance of this definition is as follows.

\[
\tau' (\theta (\kappa)) \equiv \tau' ((\kappa \nu a_0, -\kappa \nu)) \equiv \tau (\kappa \nu a_0) - \kappa \nu c' = \kappa \nu c' - \kappa \nu c' = 0
\]

This is clearly a morphism which agrees with \( \tau \) on \( B \oplus 0 \) provided we identify this with \( B \). The above shows that \( \theta (D) \subseteq \ker (\tau') \). But also, \( \theta (D) = \ker (\eta) \) from the above problem. Define

\[
\tilde{\tau}' : (B \oplus D) / \theta (D) = (B \oplus D) / \ker (\eta) \to C
\]

as

\[
\tilde{\tau}' ((b, \kappa) + \theta (D)) \equiv \tau' (b, \kappa) \equiv \tau (b) + \kappa c'
\]

This is well defined because if \((\tilde{b}, \tilde{\kappa}) - (b, \kappa) \in \theta (D)\), then this difference is in \( \ker (\tau') \) and so \( \tau' (\tilde{b}, \tilde{\kappa}) = \tau' (b, \kappa) \). Now consider the following diagram.

\[
\begin{array}{ccc}
(B \oplus D) / \theta (D) & \xrightarrow{\tilde{\eta}} & A \\
\downarrow \tilde{\tau}' & & \downarrow \sigma \\
C & \xrightarrow{i} & C
\end{array}
\]

where

\[
\tilde{\eta} : (B \oplus D) / \theta (D) = (B \oplus D) / \ker (\eta) \to A
\]

be given by \( \tilde{\eta} ((b, \kappa) + \ker (\eta)) \equiv \eta ((b, \kappa)) \equiv b + \kappa a_0 \). Then choose \( \sigma \) to make the diagram commute. That is, \( \sigma = i \circ \tilde{\tau}' \circ \tilde{\eta}^{-1} \). Verify this gives the desired extension. It is well defined. You just need to check that it agrees with \( \tau \) on \( B \). If

\[
b \in B, \tilde{\eta}^{-1} (b) \equiv (b, 0) + \theta (D)
\]

Now recall that it was shown above that \( \tau' \) extends \( \tau \). This proves the existence of the desired extension \( \sigma \).

This proves the following lemma: Let \( C = Dc_0 \) be a cyclic module over \( D \) and let \( B \) be a submodule of \( A \), also over \( D \). Suppose \( A = \text{span} (B, a_0) \) so that \( A/B \) is cyclic and suppose \( \mu A = 0 \) where \( (\mu) = \text{ann} (C) \) and let \( (\nu) \) be \( \text{ann} (B) \). Suppose that \( \tau : B \to C \) is a morphism. Then there is an extension \( \sigma : A \to C \) such that \( \sigma \) is also a morphism.
16. Let $A$ be a Noetherian torsion module, let $B$ be a sub module, and let $\tau : B \to C$ be a morphism. Here $C$ is a cyclic module $C = D c_0$ where $\text{ann}(c_0) = (\mu)$ and $\mu(A) = 0$. Then there exists a morphism $\sigma : A \to C$ which extends $\tau$. Note that here you are not given that $A/B$ is cyclic but $A$ is a Noetherian torsion module. **Hint:** Explain why there are finitely many \(\{a_1, a_2, \ldots, a_n\}\) such that $A = \text{span}(B, a_1, a_2, \ldots, a_n)$. Now explain why if $A_k = \text{span}(B, \ldots, a_k)$ then each of these is a module and $A_k/A_{k-1}$ is cyclic. Now apply the above result of the previous problem to get a succession of extensions.

17. Suppose $C$ is a cyclic sub module of $A$, a Noetherian torsion module and $(\mu) = \text{ann}(C)$. Also suppose that $\mu A = 0$. Then there exists a morphism $\sigma : A \to C$ such that $A = C \oplus \ker \sigma$

To so this, use the above proposition to get the following diagram to commute.

\[
\begin{array}{ccc}
C & \xrightarrow{\sigma} & A \\
\downarrow & & \downarrow \\
C & \xrightarrow{i} & \sigma \\
\end{array}
\]

Then consider $a = \sigma a + (a - \sigma a)$ which is something in $C$ added to something else. Explain why $\sigma^2 a = \sigma \left( \sigma^C (a) \right) = \sigma (a)$. Thus the second term is in $\ker (\sigma)$. Next suppose you have $b \in \ker (\sigma)$ and $c \in C$. Then say $c + b = 0$. Thus $\sigma c = c = 0$ when $\sigma$ is done to both sides. This gives a condition under which $C$ is a direct summand.

18. Let $M$ be a non zero torsion module for a p.i.d. $D$ and suppose that

\[
M = D z_1 + \cdots + D z_p
\]

so that it is finitely generated. Show that it is the direct sum of cyclic submodules. To do this, pick $a_1 \neq 0$ and consider $Da_1$ a cyclic submodule. Then you use the above problem to write

\[
M = Da_1 \oplus \ker \sigma
\]

for a morphism $\sigma$. Let $M_2 \equiv \ker \sigma$. It is also a torsion module for $D$. If it is not zero, do the same thing for it that was just done. The process must end because it was shown above in the chapter that $M$ is Noetherian, Proposition [72x716]9.5.4.

19. The companion matrix of the polynomial $q(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ is

\[
C = \begin{pmatrix}
0 & 0 & -a_0 \\
1 & \ddots & -a_1 \\
\vdots & \ddots & \ddots \\
0 & 1 & -a_{n-1}
\end{pmatrix}
\]

Show that the characteristic polynomial, $\det (xI - C)$ of $C$ is equal to $x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$. **Hint:** You need to take

\[
\det \begin{pmatrix}
x & 0 & a_0 \\
-1 & x & a_1 \\
\vdots & \ddots & \vdots \\
0 & -1 & x + a_{n-1}
\end{pmatrix}
\]

To do this, use induction and expand along the top row.
20. Letting $C$ be the $n \times n$ matrix as in the above where $q(x)$ is also given there, explain why
\[ Ce_1 = e_2, Ce_2 = e_3, \ldots, Ce_{n-1} = e_n. \]
Thus $e_k = C^{k-1}e_1$. Also explain why
\[ Ce_n = C^n e_1 = (-1)^k (a_0 e_1 + a_1 e_2 + \cdots + a_{n-1} e_n) = (-1)^k (a_0 e_1 + a_1 C e_1 + \cdots + a_{n-1} C^{n-1} e_1) \]
Explain why $q(C) e_1 = 0$. Now each $e_k$ is a multiple of $C$ times $e_1$. Explain why $q(C) e_k = 0$. Why does this imply that $q(C) = 0$?

21. Suppose you have a block diagonal matrix
\[ A = \begin{pmatrix} M_1 & & \\ & \ddots & \\ & & M_m \end{pmatrix} \]
Show that
\[ \det(xI - A) = \prod_{j=1}^{m} \det(xI_j - M_j), \quad I_j \text{ the appropriate size.} \]

22. Now let $L \in \mathcal{L}(V,V)$. The rational canonical form says that there is a basis for $V$ such that with respect to this basis, the matrix of $L$ is of the form
\[ B = \begin{pmatrix} M_1 & 0 \\ & \ddots \\ & 0 & M_m \end{pmatrix} \]
and each $M_j$ is also a block diagonal matrix,
\[ M_j = \begin{pmatrix} C_{j,1} & 0 \\ & \ddots \\ & 0 & C_{j,s_j} \end{pmatrix} \]
where $C_{j,a}$ is a companion matrix for a polynomial $q_{j,a}(x)$. From the above problems,
\[ \det(xI - C_{j,a}) = q_{j,a}(x), \quad q_{j,a}(C_{j,a}) = 0 \]
Explain why the characteristic polynomial for $L$ is
\[ g(x) = \det(xI - B) = \prod_{j=1}^{m} \prod_{a=1}^{s_j} \det(xI - C_{j,a}) = \prod_{j=1}^{m} \prod_{a=1}^{s_j} q_{j,a}(x) \]
Now, if you have a polynomial $g(x)$ and a block diagonal matrix
\[ D = \begin{pmatrix} C_1 & 0 \\ & \ddots \\ & 0 & C_q \end{pmatrix} \]
Explain why $g(D) =
\begin{pmatrix} g(C_1) & 0 \\ & \ddots \\ & 0 & g(C_q) \end{pmatrix}$
Now explain why $L$ satisfies its characteristic equation. This gives a general proof of the Cayley Hamilton theorem.
23. If you can find the eigenvalues exactly, then the Jordan form can be computed. Otherwise, you can forget about it. Here is a matrix.

\[
\begin{pmatrix}
1 & 2 & 0 & 0 \\
0 & 1 & 0 & 0 \\
-1 & 0 & 0 & -2 \\
1 & -1 & 1 & 3
\end{pmatrix}
\]

It has two eigenvalues, \( \lambda = 1 \), and 2. Find its Jordan form.

24. Let the field of scalars be \( \mathbb{Z}_5 \) and consider the matrix

\[
A = \begin{pmatrix}
2 & 1 & 4 \\
0 & 2 & 2 \\
0 & 0 & 0
\end{pmatrix}
\]

where the entries are considered as residue classes in \( \mathbb{Z}_5 \). Show that its eigenvalues are indeed 0, 2, 2. Now find its Jordan canonical form \( J \) and \( S \) such that \( A = SJS^{-1} \).

25. Recall that an integral domain is a commutative ring which has 1 such that if \( ab = 0 \) then one of \( a \) or \( b \) equals 0. A Euclidean domain \( D \) is one which has the property that it has a function \( \delta \), defined on the nonzero elements of \( D \) which has values in \( \{0, 1, 2, \cdots\} \) such that if \( a, b \neq 0 \) then there exist \( q, r \) such that

\[
a = bq + r, \text{ where } \delta(r) < \delta(b) \text{ or } r = 0
\]

Show that such a Euclidean domain is a principal ideal domain. Also show that an example of such a thing is the integers and also the ring of polynomials \( \mathbb{F}[x] \) for \( \mathbb{F} \) a field. Hint: Start with an ideal \( I \). Pick \( b \in I \) which has \( \delta(b) \) as small as possible. Now suppose you have \( a \in I \). Use the above Euclidean algorithm to write \( a = bq + r \) where \( \delta(r) < \delta(b) \) or else \( r = 0 \). This was done in the chapter, but do it yourself.

26. Suppose you have an \( n \times n \) matrix \( A \) in which the entries come from a commutative ring which has a multiplicative identity 1. Explain why there exists a matrix \( B \) also having entries in the ring such that \( AB = BA = I \) whenever \( \det(A) = 1 \). Explain why examples of such invertible matrices are the elementary matrices in which a multiple of a row (column) is added to another, and those which involve switching two rows (columns). Hint: You might look at determinants and the formula for the inverse of a matrix using the cofactor matrix.

27. Consider the \( m \times n \) matrices which have entries in a commutative ring \( R \). Then \( A \sim B \) means there are invertible matrices \( P, Q \) of the right size having entries in \( R \), such that

\[
A = PBQ
\]

Show that this is an equivalence relation. Note that the \( P, Q \) are different and may even be of different size. Thus this is not the usual notion of similar square matrices.

28. Suppose \( a \neq 0 \). Show that there exists an invertible \( 2 \times 2 \) matrix \( R \) such that

\[
\begin{pmatrix}
a & b \\
* & *
\end{pmatrix}
\begin{pmatrix}
d & 0 \\
* & *
\end{pmatrix}
\]

where \( d \) is a greatest common divisor of \( a \) and \( b \).

29. Let \( D \) be a principal ideal domain and let \( p \) be a noninvertible prime. Consider the principal ideal \( (p) \). Show that if \( U \) is any ideal such that \( (p) \subseteq U \subseteq D \), then either \( U = D \) or \( U = (p) \). Such an ideal is called a maximal ideal. It is an ideal, not equal to the whole ring such that no ideal can be placed between it and the whole ring. Now show that if you have a maximal ideal \( I \) in a principal ideal domain, then it must be the case that \( I = (p) \) for some prime \( p \).
30. Let \( R \) be the ring of continuous functions defined on \([0, 1]\). Here it is understood that \( f = g \) means the usual thing, that \( f(x) = g(x) \). Multiplication and addition are defined in the usual way. Pick \( x_0 \in [0, 1] \) and let \( I_{x_0} \equiv \{ f \in R : f(x_0) = 0 \}. \) Show that this is a maximal ideal of \( R \). Then show that there are no other maximal ideals. **Hint:** For the second part, let \( I \) be a maximal ideal. Show using a compactness argument and continuity of the functions that unless there exists some \( x_0 \) for which all \( f \in I \) are zero, then there exists a function in \( I \) which is never 0. Then since this is an ideal, you can show that it contains 1. Explain why this ring cannot be an integral domain.

31. Suppose you have a commutative ring which has 1 called \( R \) and suppose that \( I \) is a maximal ideal. Show that \( R/I \) is a field. **Hint:** Consider the ideal \( Rr + I \) where \( r + I \neq 0 \). Then explain why this ideal is strictly larger than \( I \) and hence equals \( R \). Thus \( 1 \in Rr + I \).

32. In the above problem about the ring of continuous functions with field of scalars \( R \), there were maximal ideals as described there. Thus \( R/I_{x_0} \) is a field. Describe the field.

33. Say you have a ring \( R \) and \( a \in R \setminus \{1\}, a \neq 0 \) which is not invertible. Explain why \( (a) \) is an ideal which does not contain 1. Show there exists a maximal ideal. **Hint:** You could let \( \mathcal{F} \) denote all ideals which do not contain 1. It is nonempty by assumption. Now partially order this by set inclusion. Consider a maximal chain. This uses the Hausdorff maximal theorem in the appendix.

34. It is always assumed that the rings used here are commutative rings and that they have a multiplicative identity 1. However, sometimes people have considered things which they have called rings \( R \) which have all the same axioms except that there is no multiplicative identity 1. However, all such things can be considered to be in a sense embedded in a real ring which has a multiplicative identity. You consider \( \mathbb{Z} \times R \) and define addition in the obvious way

\[
(k, r) + (\hat{k}, \hat{r}) \equiv \left(k + \hat{k}, r + \hat{r}\right)
\]

and multiplication as follows.

\[
(k, r) (\hat{k}, \hat{r}) \equiv \left(k\hat{k}, kr + \hat{k}r + r\hat{r}\right)
\]

Then the multiplicative identity is just \((1, 0)\). You have \((1, 0) (k, r) \equiv (k, r)\). You just have to verify the other axioms like the distributive laws and that multiplication is associative.
Chapter 10

Related Topics

This chapter is on some topics which don’t usually appear in linear algebra texts but which seem to be related to linear algebra.

10.1 The Symmetric Polynomial Theorem

First here is a definition of polynomials in many variables which have coefficients in a commutative ring. A commutative ring would be a field except you don’t know that every nonzero element has a multiplicative inverse. If you like, let these coefficients be in a field. It is still interesting. A good example of a commutative ring is the integers. In particular, every field is a commutative ring. Thus, a commutative ring satisfies the following axioms. They are just the field axioms with one omission mentioned above. You don’t have $x^{-1}$ if $x \neq 0$. We will assume that the ring has 1, the multiplicative identity.

**Axiom 10.1.1** Here are the axioms for a commutative ring.

1. $x + y = y + x$, (commutative law for addition)
2. There exists 0 such that $x + 0 = x$ for all $x$, (additive identity).
3. For each $x \in \mathbb{F}$, there exists $-x \in \mathbb{F}$ such that $x + (-x) = 0$, (existence of additive inverse).
4. $(x + y) + z = x + (y + z)$, (associative law for addition).
5. $xy = yx$, (commutative law for multiplication). You could write this as $x \times y = y \times x$.
6. $(xy)z = x(yz)$, (associative law for multiplication).
7. There exists 1 such that $1x = x$ for all $x$, (multiplicative identity).
8. $x(y + z) = xy + xz$, (distributive law).

Next is a definition of what is meant by a polynomial.

**Definition 10.1.2** Let $\mathbf{k} \equiv (k_1, k_2, \cdots, k_n)$ where each $k_i$ is a nonnegative integer. Let

$$|\mathbf{k}| \equiv \sum_{i} k_i$$

Polynomials of degree $p$ in the variables $x_1, x_2, \cdots, x_n$ are expressions of the form

$$g(x_1, x_2, \cdots, x_n) = \sum_{|\mathbf{k}| \leq p} a_{\mathbf{k}} x_1^{k_1} \cdots x_n^{k_n}$$
where each \( a_k \) is in a commutative ring. If all \( a_k = 0 \), the polynomial has no degree. Such a polynomial is said to be symmetric if whenever \( \sigma \) is a permutation of \( \{1, 2, \ldots, n\} \),

\[
g(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)}) = g(x_1, x_2, \ldots, x_n)
\]

An example of a symmetric polynomial is

\[
s_1(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i
\]

Another one is

\[
s_n(x_1, x_2, \ldots, x_n) = x_1x_2 \cdots x_n
\]

**Definition 10.1.3** The elementary symmetric polynomial \( s_k(x_1, x_2, \ldots, x_n), k = 1, \ldots, n \) is the coefficient of \((-1)^k x^{n-k}\) in the following polynomial.

\[
(x - x_1)(x - x_2) \cdots (x - x_n) = x^n - s_1x^{n-1} + s_2x^{n-2} - \cdots \pm s_n
\]

Thus

\[
s_1 = x_1 + x_2 + \cdots + x_n
\]

\[
s_2 = \sum_{i<j} x_ix_j, \quad s_3 = \sum_{i<j<k} x_ix_jx_k, \ldots, s_n = x_1x_2 \cdots x_n
\]

Then the following result is the fundamental theorem in the subject. It is the symmetric polynomial theorem. It says that these elementary symmetric polynomials raised to powers are a lot like a basis for the symmetric polynomials. This is a really remarkable result.

**Theorem 10.1.4** Let \( g(x_1, x_2, \ldots, x_n) \) be a symmetric polynomial. Then \( g(x_1, x_2, \ldots, x_n) \) equals a polynomial in the elementary symmetric polynomials.

\[
g(x_1, x_2, \ldots, x_n) = \sum_k a_k s_1^{k_1} \cdots s_n^{k_n}
\]

and the \( a_k \) in the commutative ring are unique.

**Proof:** If \( n = 1 \), it is obviously true because \( s_1 = x_1 \). Suppose the theorem is true for \( n - 1 \) and \( g(x_1, x_2, \ldots, x_n) \) has degree \( d \). Let

\[
g'(x_1, x_2, \ldots, x_{n-1}) = g(x_1, x_2, \ldots, x_{n-1}, 0)
\]

By induction, there are unique \( a_k \) such that

\[
g'(x_1, x_2, \ldots, x_{n-1}) = \sum_k a_k s_1^{k_1} \cdots s_{n-1}^{k_{n-1}}
\]

where \( s_i' \) is the corresponding symmetric polynomial which pertains to \( x_1, x_2, \ldots, x_{n-1} \). Note that

\[
s_k(x_1, x_2, \ldots, x_{n-1}, 0) = s_k'(x_1, x_2, \ldots, x_{n-1})
\]

Now consider

\[
g(x_1, x_2, \ldots, x_n) - \sum_k a_k s_1^{k_1} \cdots s_{n-1}^{k_{n-1}} = q(x_1, x_2, \ldots, x_n)
\]

is a symmetric polynomial and it equals 0 when \( x_n \) equals 0. Note that \( s_n = \pm x_1x_2 \cdots x_n \). Since \( q(x_1, x_2, \ldots, x_n) \) is symmetric, it is also 0 whenever \( x_i = 0 \). Therefore,

\[
q(x_1, x_2, \ldots, x_n) = s_n h(x_1, x_2, \ldots, x_n)
\]
and it follows that \( h (x_1, x_2, \cdots, x_n) \) is symmetric of degree no more than \( d - n \) and is uniquely determined. Thus, if \( g (x_1, x_2, \cdots, x_n) \) is symmetric of degree \( d \),
\[
g (x_1, x_2, \cdots, x_n) = \sum_{k} a_k s_1^{k_1} \cdots s_{n-1}^{k_{n-1}} + s_n h (x_1, x_2, \cdots, x_n)
\]
where \( h \) has degree no more than \( d - n \). Now apply the same argument to \( h (x_1, x_2, \cdots, x_n) \) and continue, repeatedly obtaining a sequence of symmetric polynomials \( h_i \), of strictly decreasing degree, obtaining expressions of the form
\[
g (x_1, x_2, \cdots, x_n) = \sum_{k} b_k s_1^{k_1} \cdots s_{n-1}^{k_{n-1}} s_n^{k_n} + s_n h_m (x_1, x_2, \cdots, x_n)
\]
Eventually \( h_m \) must be a constant or zero. By induction, each step in the argument yields uniqueness and so, the final sum of combinations of elementary symmetric functions is uniquely determined. 

Here is a very interesting result which I saw claimed in a paper by Steinberg and Redheffer on Lindemann’s theorem which follows from the above theorem.

**Theorem 10.1.5** Let \( \alpha_1, \cdots, \alpha_n \) be roots of the polynomial equation
\[
p (x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0
\]
where each \( a_i \) is an integer. Then any symmetric polynomial in the quantities \( a_n \alpha_1, \cdots, a_n \alpha_n \) having integer coefficients is also an integer. Also any symmetric polynomial in the quantities \( \alpha_1, \cdots, \alpha_n \) having rational coefficients is a rational number.

**Proof:** Let \( f (x_1, \cdots, x_n) \) be the symmetric polynomial. Thus
\[
f (x_1, \cdots, x_n) \in \mathbb{Z} [x_1 \cdots x_n], \text{ the polynomials having integer coefficients}
\]
From Theorem 10.1.4 it follows there are integers \( a_{k_1} \cdots a_{k_n} \) such that
\[
f (x_1, \cdots, x_n) = \sum_{k_1 + \cdots + k_n \leq m} a_{k_1} \cdots a_{k_n} p_1^{k_1} \cdots p_n^{k_n}
\]
where the \( p_i \) are the elementary symmetric polynomials defined as the coefficients of
\[
\prod_{j=1}^{n} (x - x_j)
\]
Thus
\[
f (a_n \alpha_1, \cdots, a_n \alpha_n) = \sum_{k_1 + \cdots + k_n = d} a_{k_1} \cdots a_{k_n} p_1^{k_1} (a_n \alpha_1, \cdots, a_n \alpha_n) \cdots p_n^{k_n} (a_n \alpha_1, \cdots, a_n \alpha_n)
\]
Now the given polynomial \( p (x) \) is of the form
\[
a_n \prod_{j=1}^{n} (x - \alpha_j) \equiv a_n \left( \sum_{k=0}^{n} p_k (\alpha_1, \cdots, \alpha_n) x^{n-k} \right)
\]
\[
= a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0
\]
and so
\[
a_n p_k (\alpha_1, \cdots, \alpha_n) = a_n a_{n-k}
\]
Note that
\[
a^n_k p_k (\alpha_1, \cdots, \alpha_n) = p_k (a_n \alpha_1, \cdots, a_n \alpha_n),
\]
\[
a_n a_{n-k} = a_n p_k (\alpha_1, \cdots, \alpha_n) = \frac{a_n}{a^n_k} p_k (a_n \alpha_1, \cdots, a_n \alpha_n)
\]
\[
p_k (a_n \alpha_1, \cdots, a_n \alpha_n) = a^n_k a_{n-k}
\]
Therefore,
\[
f (a_n \alpha_1, \cdots, a_n \alpha_n) = \sum_{k_1 + \cdots + k_n = d} a_{k_1 \cdots k_n} p_{k_1}^{k_1} \cdots p_n^{k_n} (a_n \alpha_1, \cdots, a_n \alpha_n)
\]
and each \( p_k (a_n \alpha_1, \cdots, a_n \alpha_n) \) is an integer. Thus \( f (a_n \alpha_1, \cdots, a_n \alpha_n) \) is an integer as claimed. From this, it is obvious that \( f (\alpha_1, \cdots, \alpha_n) \) is rational. Indeed,
\[
f (\alpha_1, \cdots, \alpha_n) = \sum_{k_1 + \cdots + k_n = d} a_{k_1 \cdots k_n} p_{k_1}^{k_1} \cdots p_n^{k_n} (\alpha_1, \cdots, \alpha_n)
\]
Now multiply both sides by \( a_n a_2^3 a_3^4 \cdots a^n_n \), an integer. Then
\[
a_n a_2^3 a_3^4 \cdots a^n_n f (\alpha_1, \cdots, \alpha_n) = \sum_{k_1 + \cdots + k_n = d} a_{k_1 \cdots k_n} p_{k_1}^{k_1} \cdots p_n^{k_n} (a_n \alpha_1, \cdots, a_n \alpha_n)
\]
with the right side an integer. Thus \( f (\alpha_1, \cdots, \alpha_n) \) is rational. \[\square\]

### 10.2 Transcendental Numbers

Most numbers are like this. Here the algebraic numbers are those which are roots of a polynomial equation having rational numbers as coefficients. By the fundamental theorem of algebra, all these numbers are in \( \mathbb{C} \). There are only countably many of these algebraic numbers, (Problem 14 on Page 218). Therefore, most numbers are transcendental. Nevertheless, it is very hard to prove that this or that number is transcendental. Probably the most famous theorem about this is the Lindemann-Weierstrass theorem.

**Theorem 10.2.1** Let the \( \alpha_i \) be distinct nonzero algebraic numbers and let the \( a_i \) be nonzero algebraic numbers. Then
\[
\sum_{i=1}^{n} a_i e^{\alpha_i} \neq 0
\]

I am following the interesting Wikepedia article on this subject. You can also look at the book by Baker [12], Transcendental Number Theory, Cambridge University Press. There are also many other treatments which you can find on the web including an interesting article by Steinberg and Redheffer which appeared in about 1950.

The proof makes use of the following identity. For \( f (x) \) a polynomial,
\[
I (s) \equiv \int_{0}^{s} e^{s-x} f (x) \, dx = e^{s} \sum_{j=0}^{\deg(f)} f^{(j)} (0) - \sum_{j=0}^{\deg(f)} f^{(j)} (s).
\]
(10.1)

where \( f^{(j)} \) denotes the \( j \)th derivative. In this formula, \( s \in \mathbb{C} \) and the integral is defined in the natural way as
\[
\int_{0}^{1} s f (ts) e^{s-ts} \, dt
\]
(10.2)
The identity follows from integration by parts.

\[
\int_0^1 sf(ts)e^{-ts}dt = se^s\int_0^1 f(ts)e^{-ts}dt
\]

\[
= se^s\left[-\frac{e^{-ts}}{s}f(ts)\bigg|_0^1 + \int_0^1 e^{-ts}s f'(st)dt\right]
\]

\[
= se^s\left[\frac{1}{s}f(s) - \frac{e^{-s}}{s}f(0) + \int_0^1 e^{-ts}f'(st)dt\right]
\]

\[
f(0) - e^s f(s) + \int_0^1 se^{s-ts}f'(st)dt
\]

\[
= f(0) - f(s)e^s + \int_0^s e^{s-x}f'(x)dx
\]

Continuing this way establishes the identity since at the right end looks just like what we started with except with a derivative on the \(f\).

**Lemma 10.2.2** If \(K\) and \(c\) are nonzero integers, and \(\beta_1, \cdots, \beta_m\) are the roots of a single polynomial with integer coefficients,

\[Q(x) = vx^m + \cdots + u\]

where \(v, u \neq 0\), then

\[K + c(e^{\beta_1} + \cdots + e^{\beta_m}) \neq 0.\]

Letting

\[f(x) = \frac{v(m+1)p Q^p(x) x^{p-1}}{(p-1)!}\]

and \(I(s)\) be defined in terms of \(f(x)\) as above, it follows,

\[
\lim_{p \to \infty} \sum_{i=1}^m I(\beta_i) = 0
\]

and

\[
\sum_{j=0}^n f^{(j)}(0) = v^{p(m+1)}a^p + m_1(p)p
\]

\[
\sum_{i=1}^m \sum_{j=0}^n f^{(j)}(\beta_i) = m_2(p)p
\]

where \(m_i(p)\) is some integer.

**Proof:** Let \(p\) be a large prime number. Then consider the polynomial \(f(x)\) of degree \(n \equiv pm + p - 1,\)

\[f(x) = \frac{v(m+1)p Q^p(x) x^{p-1}}{(p-1)!}\]

From [12.1],

\[
c \sum_{i=1}^m I(\beta_i) = c \sum_{i=1}^m \left(e^{\beta_i} \sum_{j=0}^n f^{(j)}(0) - \sum_{j=0}^n f^{(j)}(\beta_i)\right)
\]

\[
= \left(K + c \sum_{i=1}^m e^{\beta_i}\right) \sum_{j=0}^n f^{(j)}(0) - K \sum_{j=0}^n f^{(j)}(0) - c \sum_{i=1}^m \sum_{j=0}^n f^{(j)}(\beta_i)
\]

Claim 1: \(\lim_{p \to \infty} c \sum_{i=1}^m I(\beta_i) = 0.\)
**Proof:** This follows right away from the definition of \( I (\beta_j) \) and the definition of \( f (x) \).

\[
|I (\beta_j)| \leq \int_0^1 |\beta_j f (t \beta_j) e^{t \beta_j} | \, dt \\
\leq \int_0^1 |v^{(m-1)p} Q (t \beta_j)^p \beta_j^{p-1} | \beta_j^{p-1} \frac{dt}{(p-1)!}
\]

which clearly converges to 0. This proves the claim.

The next thing to consider is the term on the end in 10.3,

\[
K \sum_{j=0}^n f^{(j)} (0) + c \sum_{i=1}^m \sum_{j=0}^n f^{(j)} (\beta_i)
\]

The idea is to show that for large enough \( p \) it is always a nonzero integer. When this is done, it can’t happen that \( K + c \sum_{i=1}^m e^{\beta_i} = 0 \) because if this were so, you would have a very small number equal to an integer.

\[
f (x) = \frac{v^{(m+1)p} (vx^m + \cdots + u)^p x^p}{(p-1)!}
\]

Then \( f^j (0) = 0 \) unless \( j \geq p-1 \) because otherwise, that \( x^{p-1} \) term will result in some \( x^r, r > 0 \) and everything is zero when you plug in \( x = 0 \). Now say \( j = p-1 \). Then it is clear that

\[
f^{(p-1)} (0) = u^p v^{(m+1)p}
\]

So what if \( j > p-1 \)? Then by Liebniz formula,

\[
f^j (x) = \binom{j}{p-1} v^{(m+1)p} \frac{d}{dx^{j-(p-1)}} [(vx^m + \cdots + u)^p] + Stuff
\]

where the \( Stuff \) equals 0 when \( x = 0 \). Thus \( f^j (0) = pm_j \) where \( m_j \) is some integer depending on the integer coefficients of the polynomial \( Q (x) \). Therefore,

\[
\sum_{j=0}^n f^{(j)} (0) = v^{(m+1)p} u^p + m (p) p
\]

where \( m (p) \) is some integer.

Now consider the other sum in 10.4,

\[
c \sum_{i=1}^m \sum_{j=0}^n f^{(j)} (\beta_i)
\]

Using the formula for \( f (x) \) and that the \( \beta_i \) are roots,

\[
f (x) = \frac{v^{(m+1)p} ((x-\beta_1) (x-\beta_2) \cdots (x-\beta_m))^p x^p}{(p-1)!}
\]

it follows that for \( j < p, \ f^{(j)} (\beta_i) = 0 \). This is because for such derivatives, each term will have that product of the \( (x-\beta_i) \) in it.

To get something non zero, the nonzero terms must involve at least \( p \) derivatives of the expression

\[
((x-\beta_1) (x-\beta_2) \cdots (x-\beta_m))^p
\]

since otherwise, when evaluated at any \( \beta_k \) the result would be 0.
Now say \( j \geq p \). Then by Liebniz formula, \( f^j (x) \) is of the form

\[
\frac{v^p^{(m+1)p}}{(p-1)!} \sum_{r=0}^{p(m+1)-2p+1} \binom{j}{r} \frac{d}{dx^r} \left( ((x - \beta_1) (x - \beta_2) \cdots (x - \beta_m))^p \right) \frac{d}{dx^{p-r}} x^{p-1}
\]

\[
= \frac{v^p^{(m+1)-2p+1}}{(p-1)!} \sum_{r=0}^{p(m+1)-2p+1} \binom{j}{r} \frac{d}{dx^r} \left( ((vx - v \beta_1) (vx - v \beta_2) \cdots (vx - v \beta_m))^p \right) \frac{d}{dx^{p-r}} (vx)^{p-1}
\]

Note that for \( r \) too small, the term will be zero when evaluated at any of the \( \beta_i \). You only get something nonzero if \( r \geq p \) and so there will be a \( p! \) produced which will cancel with the \( (p-1)! \) to yield an extra \( p \).

Now if you do the computations using the product rule and then replace \( x \) with \( \beta_i \) and sum these over all \( v \beta_i \), you will get a symmetric polynomial in the quantities \( \{ v \beta_1, \cdots, v \beta_m \} \) and by Theorem \( 1.1.8 \) this is an integer. To see this is symmetric note that switching \( v \beta_a, v \beta_b \) in

\[
\frac{d}{dx^r} \left( (vx - v \beta_1) (vx - v \beta_2) \cdots (vx - v \beta_m) \right)
\]

does not change anything. The other term is just \( v^{p-1} (p-1) (p-2) \cdots (p - j + r) x^{p-j+r-1} \) or zero if \( j - r > p - 1 \). It follows that when adding these over \( i \),

\[
c \sum_{i=1}^{m} \sum_{j=0}^{n} f^{(j)} (\beta_i) = L (p) p
\]

where \( L (p) \) is some integer. Therefore, \( K u^{p(m+1)} + m (p) p + L (p) p \equiv K u^{p(m+1)} u^p + M (p) p \) for some integer \( M (p) \). Summarizing, it follows

\[
c \sum_{i=1}^{m} I (\beta_i) = \left( K + c \sum_{i=1}^{m} e^{\beta_i} \right) \sum_{j=0}^{n} f^{(j)} (0) + K u^{p(m+1)} u^p + M (p) p
\]

where the left side is very small whenever \( p \) is large enough. Let \( p \) be larger than \( \max (K, v, u) \).

Since \( p \) is prime, it follows that it cannot divide \( K u^{p(m+1)} u^p \) and so the last two terms must sum to a nonzero integer and so the equation \( I \prod \) cannot hold unless

\[
K + c \sum_{i=1}^{m} e^{\beta_i} \neq 0 \quad \blacksquare
\]

Note that this shows \( \pi \) is irrational. If \( \pi = k/m \) where \( k, m \) are integers, then both \( i \pi \) and \( -i \pi \) are roots of the polynomial with integer coefficients,

\[
m^2 x^2 + k^2
\]

which would require, from what was just shown that

\[
0 \neq 2 + e^{i \pi} + e^{-i \pi}
\]

which is not the case since the sum on the right equals 0.

The following corollary follows from this.

**Corollary 10.2.3** Let \( K \) and \( c_i \) for \( i = 1, \cdots, n \) be nonzero integers. For each \( k \) between 1 and \( n \) let \( \{ \beta (k) \}_{i=1}^{m_k} \) be the roots of a polynomial with integer coefficients,

\[
Q_k (x) \equiv v_k x^{m_k} + \cdots + u_k
\]
where $v_k, u_k \neq 0$. Then

$$K + c_1 \left( \sum_{j=1}^{m_1} e^{\beta(1)}_j \right) + c_2 \left( \sum_{j=1}^{m_2} e^{\beta(2)}_j \right) + \cdots + c_n \left( \sum_{j=1}^{m_n} e^{\beta(n)}_j \right) \neq 0.$$  

Proof: Defining $f_k(x)$ and $I_k(s)$ as in Lemma 10.2.2, it follows from Lemma 10.2.2 that for each $k = 1, \cdots, n$,

$$c_k \sum_{i=1}^{m_k} I_k(\beta(k)_i) = \left( K_k + c_k \sum_{i=1}^{m_k} e^{\beta(k)_i} \right) \left( v^{(m_k+1)}_k u^p_k + N_k p \right)$$

$$-K_k \left( v^{(m_k+1)}_k u^p_k + N_k p \right) - c_k N'_k p$$

and so

$$c_k \sum_{i=1}^{m_k} I_k(\beta(k)_i) = \left( K_k + c_k \sum_{i=1}^{m_k} e^{\beta(k)_i} \right) \left( v^{(m_k+1)}_k u^p_k + N_k p \right)$$

$$-K_k v^{(m_k+1)}_k u^p_k + M_k p$$

for some integer $M_k$. By multiplying each $Q_k(x)$ by a suitable constant depending on $k$, it can be assumed without loss of generality that all the $v^{m_k+1}_k u_k$ are equal to a constant integer $U$. Then the above equals

$$c_k \sum_{i=1}^{m_k} I_k(\beta(k)_i) = \left( K_k + c_k \sum_{i=1}^{m_k} e^{\beta(k)_i} \right) \left( U^p + N_k p \right)$$

$$-K_k U^p + M_k p$$

Adding these for all $k$ gives

$$\sum_{k=1}^{n} c_k \sum_{i=1}^{m_k} I_k(\beta(k)_i) = U^p \left( K + \sum_{k=1}^{n} c_k \sum_{i=1}^{m_k} e^{\beta(k)_i} \right) - K U^p + M_p$$

$$+ \sum_{k=1}^{n} N_k p \left( K_k + c_k \sum_{i=1}^{m_k} e^{\beta(k)_i} \right)$$

(10.6)

For large $p$ it follows from Lemma 10.2.2 that the left side is very small. If

$$K + \sum_{k=1}^{n} c_k \sum_{i=1}^{m_k} e^{\beta(k)_i} = 0$$

then $\sum_{k=1}^{n} c_k \sum_{i=1}^{m_k} e^{\beta(k)_i}$ is an integer and so the last term in (10.6) is an integer times $p$. Thus for large $p$ it reduces to

$$\text{small number} = -K U^p + Ip$$

where $I$ is an integer. Picking prime $p > \max (U, K)$ it follows $-K U^p + Ip$ is a nonzero integer and this contradicts the left side being a small number less than 1 in absolute value. ■

Next is an even more interesting Lemma which follows from the above corollary.
Lemma 10.2.4 If \( b_0, b_1, \ldots, b_n \) are non zero integers, and \( \gamma_1, \ldots, \gamma_n \) are distinct algebraic numbers, then
\[
b_0 e^{\gamma_0} + b_1 e^{\gamma_1} + \cdots + b_n e^{\gamma_n} \neq 0
\]

Proof: Assume
\[
b_0 e^{\gamma_0} + b_1 e^{\gamma_1} + \cdots + b_n e^{\gamma_n} = 0 \tag{10.7}
\]
Divide by \( e^{\gamma_0} \) and letting \( K = b_0 \),
\[
K + b_1 e^{\alpha(1)} + \cdots + b_n e^{\alpha(n)} = 0 \tag{10.8}
\]
where \( \alpha(k) = \gamma_k - \gamma_0 \). These are still distinct algebraic numbers none of which is 0 thanks to Theorem 10.3.20. Therefore, \( \alpha(k) \) is a root of a polynomial
\[
v_k x^{m_k} + \cdots + u_k \tag{10.9}
\]
having integer coefficients, \( v_k, u_k \neq 0 \). Recall algebraic numbers were defined as roots of polynomial equations having rational coefficients. Just multiply by the denominators to get one with integer coefficients. Let the roots of this polynomial equation be
\[
\{ \alpha(k)_1, \ldots, \alpha(k)_{m_k} \}
\]
and suppose they are listed in such a way that \( \alpha(k)_1 = \alpha(k) \). Letting \( i_k \) be an integer in \( \{1, \ldots, m_k\} \) it follows from the assumption in Lemma 10.2.4 that
\[
\prod_{(i_1, \ldots, i_n) \atop i_k \in \{1, \ldots, m_k\}} \left( K + b_1 e^{\alpha(1)i_1} + b_2 e^{\alpha(2)i_2} + \cdots + b_n e^{\alpha(n)i_n} \right) = 0 \tag{10.10}
\]
This is because one of the factors is the one occurring in Lemma 10.2.4 when \( i_k = 1 \) for every \( k \). The product is taken over all distinct ordered lists \( (i_1, \ldots, i_n) \) where \( i_k \) is as indicated. Expand this possibly huge product. This will yield something like the following.
\[
K' + c_1 \left( e^{\beta(1)i_1} + \cdots + e^{\beta(1)i_1} \right) + c_2 \left( e^{\beta(2)i_2} + \cdots + e^{\beta(2)i_2} \right) + \cdots +
\]
\[
c_N \left( e^{\beta(N)i_1} + \cdots + e^{\beta(N)i_1} \right) = 0 \tag{10.11}
\]
These integers \( c_j \) come from products of the \( b_i \) and \( K \). The \( \beta(i)_j \) are the distinct exponents which result. Note that a typical term in this product would be something like
\[
\text{integer term } \frac{a(k_{1})_{i_1} + a(k_{2})_{i_2} + \cdots + a(k_{n-p})_{i_{n-p}}}{K^p b_{k_1} \cdots b_{k_{n-p}} e^{\beta(i)_r}}
\]
the \( k_r \) possibly not distinct and each \( i_k \in \{1, \ldots, m_k\} \). A given term in the sum of the integer expression corresponds to such a choice of \( \{b_{k_1}, \ldots, b_{k_{n-p}}\} \) which leads to \( K^p b_{k_1} \cdots b_{k_{n-p}} \) times a sum of exponentials like those just described. Since the product in Lemma 10.2.4 is taken over all choices \( i_k \in \{1, \ldots, m_k\} \), it follows that if you switch \( \alpha(r)_1 \) and \( \alpha(r)_2 \), two of the roots of the polynomial
\[
v_r x^{m_r} + \cdots + u_r
\]
mentioned above, the result in Lemma 10.2.4 would be the same except for permuting the
\[
\beta(s)_1, \beta(s)_2, \ldots, \beta(s)_{\mu(s)}
\]
Thus a symmetric polynomial in
\[
\beta(s)_1, \beta(s)_2, \ldots, \beta(s)_{\mu(s)}
\]
is also a symmetric polynomial in the \( \alpha(k)_1, \alpha(k)_2, \ldots, \alpha(k)_{m_k} \) for each \( k \). Thus for a given \( r, \beta(r)_1, \ldots, \beta(r)_{\mu(r)} \) are roots of the polynomial

\[
(x - \beta(r)_1)(x - \beta(r)_2) \cdots (x - \beta(r)_{\mu(r)})
\]

whose coefficients are symmetric polynomials in the \( \beta(r)_j \) which is a symmetric polynomial in the \( \alpha(k)_j, j = 1, \ldots, m_k \) for each \( k \). Letting \( g \) be one of these symmetric polynomials, a coefficient of the above polynomial, and writing it in terms of the \( \alpha(k)_j \) you would have

\[
\sum_{l_1, \ldots, l_n} A_{l_1, \ldots, l_n} \alpha(n)_1^{l_1} \alpha(n)_2^{l_2} \cdots \alpha(n)_{m_n}^{l_n}
\]

where \( A_{l_1, \ldots, l_n} \) is a symmetric polynomial in \( \alpha(k)_j, j = 1, \ldots, m_k \) for each \( k \leq n - 1 \). (It is desired to show \( g \) is rational.) These coefficients are in the field \( \mathbb{Q}[A(1), \ldots, A(n-1)] \) where \( A(k) \) denotes

\[
\{ \alpha(k)_1, \ldots, \alpha(k)_{m_k} \}
\]

and so from Theorem 10.2.5, the above symmetric polynomial is of the form

\[
\sum_{(k_1, \ldots, k_{m_n})} B_{k_1, \ldots, k_{m_n}} p_1^{k_1}(\alpha(n)_1, \ldots, \alpha(n)_{m_n}) \cdots p_{m_n}^{k_{m_n}}(\alpha(n)_1, \ldots, \alpha(n)_{m_n})
\]

where \( B_{k_1, \ldots, k_{m_n}} \) is a symmetric polynomial in \( \alpha(k)_j, j = 1, \ldots, m_k \) for each \( k \leq n - 1 \). Now do for each \( B_{k_1, \ldots, k_{m_n}} \) what was just done for \( g \) featuring this time

\[
\{ \alpha(n-1)_1, \ldots, \alpha(n-1)_{m_{n-1}} \}
\]

and continuing this way, it must be the case that eventually you have a sum of integer multiples of products of elementary symmetric polynomials in \( \alpha(k)_j, j = 1, \ldots, m_k \) for each \( k \leq n \). By Theorem 10.2.4, these are rational numbers. Therefore, each such \( g \) is a rational number and so the \( \beta(r)_j \) are algebraic roots of a polynomial having rational coefficients, hence also roots of one which has integer coefficients. Now 10.2.4 contradicts Corollary 10.2.6. \( \square \)

Note this lemma is sufficient to prove Lindemann’s theorem that \( \pi \) is transcendental. Here is why. If \( \pi \) is algebraic, then so is \( i\pi \) and so from this lemma, \( e^0 + e^{i\pi} \neq 0 \) but this is not the case because \( e^{i\pi} = -1 \).

The next theorem is the main result, the Lindemann-Weierstrass theorem.

**Theorem 10.2.5** Suppose \( a(1), \ldots, a(n) \) are nonzero algebraic numbers and suppose \( \alpha(1), \ldots, \alpha(n) \) are distinct algebraic numbers. Then

\[
a(1)e^{\alpha(1)} + a(2)e^{\alpha(2)} + \cdots + a(n)e^{\alpha(n)} \neq 0
\]

**Proof:** Suppose \( a(j) \equiv a(j)_1 \) is a root of the polynomial

\[
v_j x^{m_j} + \cdots + u_j
\]

where \( v_j, u_j \neq 0 \). Let the roots of this polynomial be \( a(j)_1, \ldots, a(j)_{m_j} \). Suppose to the contrary that

\[
a(1)_1 e^{\alpha(1)} + a(2)_1 e^{\alpha(2)} + \cdots + a(n)_1 e^{\alpha(n)} = 0
\]

Then consider the big product

\[
\prod_{(i_1, \ldots, i_n) \in \{1, \ldots, m_k\}} \left( a(1)_{i_1} e^{\alpha(1)} + a(2)_{i_2} e^{\alpha(2)} + \cdots + a(n)_{i_n} e^{\alpha(n)} \right) \tag{10.12}
\]
the product taken over all ordered lists \((i_1, \cdots, i_n)\). This product equals

\[
0 = b_1 e^{\beta(1)} + b_2 e^{\beta(2)} + \cdots + b_N e^{\beta(N)}
\]

(10.13)

where the \(\beta(j)\) are the distinct exponents which result. The \(\beta(i)\) are clearly algebraic because they are the sum of the \(\alpha(i)\). Since the product in \(\text{Eq. (10.14)}\) is taken for all ordered lists as described above, it follows that for a given \(k\), if \(\alpha(k)\) is switched with \(\alpha(k)\), that is, two of the roots of \(v_k a_{mk} + \cdots + u_k\) are switched, then the product is unchanged and so \(\text{Eq. (10.13)}\) is also unchanged. Thus each \(b_k\) is a symmetric polynomial in the \(a(k)_j\), \(j = 1, \cdots, m_k\) for each \(k\). It follows

\[
b_k = \sum_{(j_1, \cdots, j_{m_k})} A_{j_1, \cdots, j_{m_k}} a(n)^{j_1}_{1} \cdots a(n)^{j_{m_k}}_{m_k}
\]

and this is symmetric in the \(\{a(n)_1, \cdots, a(n)_{m_k}\}\) the coefficients \(A_{j_1, \cdots, j_{m_k}}\) being in the field \(\mathbb{Q}[A(1), \cdots, A(n-1)]\) where \(A(k)\) denotes

\[
a(k)_1, \cdots, a(k)_{m_k}
\]

and so from Theorem \(\text{10.16}^\star\),

\[
b_k = \sum_{(j_1, \cdots, j_{m_k})} B_{j_1, \cdots, j_{m_k}} p_1^{j_1}(a(n)_1 \cdots a(n)_{m_k}) \cdots p_{m_k}^{j_{m_k}}(a(n)_1 \cdots a(n)_{m_k})
\]

where the \(B_{j_1, \cdots, j_{m_k}}\) are symmetric in \(\{a(k)_j\}_{j=1}^{m_k}\) for each \(k \leq n - 1\). Now doing to \(B_{j_1, \cdots, j_{m_k}}\) what was just done to \(b_k\) and continuing this way, it follows \(b_k\) is a finite sum of integers times elementary polynomials in the various \(\{a(k)_j\}_{j=1}^{m_k}\) for \(k \leq n\). By Theorem \(\text{10.16}^\star\) this is a rational number. Thus \(b_k\) is a rational number. Multiplying by the product of all the denominators, it follows there exist integers \(c_i\) such that

\[
0 = c_1 e^{\beta(1)} + c_2 e^{\beta(2)} + \cdots + c_N e^{\beta(N)}
\]

which contradicts Lemma \(\text{10.2.8}^\star\). \(\square\)

This theorem is sufficient to show \(e\) is transcendental. If it were algebraic, then

\[
e e^{-1} + (-1) e^0 \neq 0
\]

but this is not the case. If \(a \neq 1\) is algebraic, then \(\ln(a)\) is transcendental. To see this, note that

\[
1 e^{\ln(a)} + (-1) a e^0 = 0
\]

which cannot happen if \(\ln(a)\) is algebraic according to the above theorem. If \(a\) is algebraic and \(\sin(a) \neq 0\), then \(\sin(a)\) is transcendental because

\[
\frac{1}{2i} e^{i a} - \frac{1}{2i} e^{-i a} + (-1) \sin(a) e^0 = 0
\]

which cannot occur if \(\sin(a)\) is algebraic. There are doubtless other examples of numbers which are transcendental by this amazing theorem.

### 10.3 The Fundamental Theorem Of Algebra

This is devoted to a mostly algebraic proof of the fundamental theorem of algebra. It depends on the interesting results about symmetric polynomials which are presented above. I found it on the Wikipedia article about the fundamental theorem of algebra. You google “fundamental theorem of algebra” and go to the Wikipedia article. It gives several other proofs in addition to this one. According to this article, the first completely correct proof of this major theorem is due to Argand in 1806. Gauss and others did it earlier but their arguments had gaps in them.

You can’t completely escape analysis when you prove this theorem. The necessary analysis is in the following lemma. Recall that \(\mathbb{A}\) is the field of algebraic numbers.
Lemma 10.3.1 Let \( \mathbb{A} \) denote the algebraic numbers. Suppose \( p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \) where \( n \) is odd and the coefficients are rational. Then \( p(x) \) has a root in \( \mathbb{A} \).

**Proof:** This follows from the intermediate value theorem from calculus. There is a real root and by definition it is in \( \mathbb{A} \). \( \blacksquare \)

Next is an algebraic consideration. First recall some notation.

\[
\prod_{i=1}^{m} a_i = a_1a_2 \cdots a_m
\]

Recall a polynomial in \( \{z_1, \cdots, z_n\} \) is symmetric only if it can be written as a sum of elementary symmetric polynomials raised to various powers multiplied by constants. This follows from Theorem 10.1.4, the theorem on symmetric polynomials.

The following is the main part of the theorem. In fact this is one version of the fundamental theorem of algebra which people studied earlier in the 1700’s.

Lemma 10.3.2 Let \( p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \) be a polynomial with rational coefficients. Then it has a root in \( \mathbb{A} \) where \( \mathbb{A} \) denotes the algebraic numbers.

**Proof:** It is possible to write

\[
n = 2^km
\]

where \( m \) is odd. If \( n \) is odd, \( k = 0 \). If \( n \) is even, keep dividing by 2 until you are left with an odd number. If \( k = 0 \) so that \( n \) is odd, it follows from Lemma 10.3.1 that \( p(x) \) has a real, hence algebraic root which is in \( \mathbb{A} \). This follows from the definition of algebraic numbers. Thus the case where \( k = 0 \) is done.

The proof will be by induction on \( k \). Suppose then that it works for \( n = 2^lm \) where \( m \) is odd and \( l \leq k-1 \) and let \( n = 2^km \) where \( m \) is odd. Let \( \{z_1, \cdots, z_n\} \) be the roots of the polynomial in a splitting field, the existence of this field being given earlier. Then

\[
p(x) = \prod_{j=1}^{n} (x - z_j) = \sum_{k=0}^{n} (-1)^k p_k(z_1, \cdots, z_n) x^k
\]

(10.14)

where \( p_k \) is the \( k^{th} \) elementary symmetric polynomial. Note this shows

\[
a_{n-k} = p_k(-1)^k.
\]

(10.15)

The thing to take away from here is that the elementary symmetric polynomials \( p_k \) are rational numbers.

There is another polynomial which has coefficients which are sums of rational numbers times the \( p_k \)'s raised to various powers and it is

\[
q_t(x) = \prod_{1 \leq i < j \leq n} (x - (z_i + z_j + tz_iz_j)), t \in \mathbb{Q}
\]

I need to verify this is really the case for \( q_t(x) \). When you switch any two of the \( z_i \) in \( q_t(x) \) the polynomial does not change. For example, let \( n = 3 \) when \( q_0(x) \) is

\[
(x - (z_1 + z_2 + t_1z_1z_2))(x - (z_1 + z_3 + t_1z_1z_3))(x - (z_2 + z_3 + t_1z_2z_3))
\]

and you can observe the assertion about the polynomial is true when you switch two different \( z_i \). Thus the coefficients of \( q_t(x) \) must be symmetric polynomials in the \( z_i \) with rational coefficients. Hence by Proposition 10.1.5 these coefficients are polynomials with rational coefficients in terms of the elementary symmetric polynomials \( p_k, k = 1, \cdots, n \). Thus by 10.1.7 the coefficients of \( q_t(x) \) are polynomials in terms of the \( a_k \) of the original polynomial. Recall these were all rational. It follows, and this is what was wanted, that \( q_t(x) \) has all rational coefficients.
10.3. THE FUNDAMENTAL THEOREM OF ALGEBRA

Note that the degree of \( q_t(x) \) is \( \binom{n}{2} \) because there are this number of ways to pick \( i < j \) out of \( \{1, \ldots, n\} \). Now

\[
\binom{n}{2} = \frac{n(n-1)}{2} = 2^{k-1}m(2^km-1)
\]

and so by induction, for each \( t \in \mathbb{Q} \), \( q_t(x) \) has a root which is in \( \mathbb{A} \).

There must exist \( s \neq t \) such that for a single pair of indices \( i, j \), with \( i < j \),

\[
(z_i + z_j + tz_i z_j), (z_i + z_j + s z_i z_j)
\]

are both in \( \mathbb{A} \). Here is why. Let \( A(i, j) \) denote those \( t \in \mathbb{R} \) such that \( (z_i + z_j + t z_i z_j) \) is complex. It was just shown that every \( t \in \mathbb{Q} \) must be in some \( A(i, j) \). There are infinitely many \( t \in \mathbb{Q} \) and so some \( A(i, j) \) contains two of them.

Now for that \( t, s \),

\[
z_i + z_j + t z_i z_j = a \quad \text{and} \quad z_i + z_j + s z_i z_j = b
\]

where \( t \neq s \) and so by Cramer’s rule,

\[
\begin{vmatrix}
  a & t \\
  b & s \\
  1 & t \\
  1 & s
\end{vmatrix} \in \mathbb{A}
\]

and also

\[
\begin{vmatrix}
  1 & a \\
  1 & b \\
  1 & t \\
  1 & s
\end{vmatrix} \in \mathbb{A}
\]

At this point, note that \( z_i, z_j \) are both solutions to the equation

\[
x^2 - (z_1 + z_2) x + z_1 z_2 = 0,
\]

which from the above has roots in \( \mathbb{A} \) by the quadratic formula. Thus the original polynomial has a root in \( \mathbb{A} \). ■

With this lemma, it is easy to prove the fundamental theorem of algebra. The difference between the lemma and this theorem is that in the theorem, the coefficients are only assumed to be in \( \mathbb{A} \). What this means is that if you have any polynomial with coefficients in \( \mathbb{A} \), it is not irreducible with respect to this field. Hence the field extension is the same field. Another way to say this is that for every polynomial with coefficients in \( \mathbb{A} \), there exists a factorization into linear factors or in other words a splitting field for a polynomial with coefficients in \( \mathbb{A} \) is \( \mathbb{A} \).

**Theorem 10.3.3** Let \( p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \) be any polynomial having coefficients in \( \mathbb{A} \), \( n \geq 1, a_n \neq 0 \). Then it has a root in \( \mathbb{A} \). Furthermore, there exist complex numbers \( z_1, \ldots, z_n \) such that

\[
p(x) = a_n \prod_{k=1}^{n} (x - z_k)
\]
CHAPTER 10. RELATED TOPICS

Proof: First suppose \( a_n = 1 \). Consider the polynomial
\[ q(x) \equiv p(x) p(x) \]
this is a polynomial and it has coefficients in \( A \). This is because it equals
\[
(x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0) \cdot \\
(x^n + \overline{a_{n-1}}x^{n-1} + \cdots + \overline{a_1}x + \overline{a_0})
\]
The \( x^{j+k} \) term of the above product is of the form
\[ a_k x^k \overline{a_j} x^j + \overline{a_k} a_j x^j = x^{k+j} (a_k \overline{a_j} + \overline{a_k} a_j) \]
and
\[ a_k \overline{a_j} + \overline{a_k} a_j = a_k \overline{a_j} + \overline{a_k} a_j \]
so it is of the form of a number in \( A \) added to its conjugate. Hence \( q(x) \) has coefficients in \( A \) as claimed. Therefore, by Lemma 10.3.2 it has a root \( z \in A \). Hence \( p(z) = 0 \). Thus \( p(x) \) has a root in \( A \).

Next suppose \( a_n \neq 0 \). Then simply divide by it and get a polynomial in which \( a_n = 1 \). Denote this modified polynomial as \( q(x) \). Then by what was just shown and the Euclidean algorithm, there exists \( z_1 \in A \) such that
\[ q(x) = (x - z_1) q_1(x) \]
where \( q_1(x) \) has coefficients in \( A \). Now do the same thing for \( q_1(x) \) to obtain
\[ q(x) = (x - z_1)(x - z_2) q_2(x) \]
and continue this way. Thus
\[ p(x/a_n) = \prod_{j=1}^{n} (x - z_j) \]

Obviously this is a harder proof than the other proof of the fundamental theorem of algebra presented earlier. However, this shows that \( A \) is algebraically complete rather than \( \mathbb{C} \) which is totally impossible to show using the earlier analytical proof. To get the case of \( \mathbb{C} \) which is the usual fundamental theorem of algebra, just replace \( \mathbb{Q} \) in the above with \( \mathbb{R} \).

Recall from Problem 10 on Page 72 that \( A \) is countable and so this is also the case for \( A \). Thus this gives an algebraically complete field which is countable and so very different than \( \mathbb{C} \). Note also that \( A \) is not complete in the sense that Cauchy sequences might not converge.

10.4 More On Algebraic Field Extensions

The next few sections have to do with fields and field extensions. There are many linear algebra techniques which are used in this discussion and it seems to me to be very interesting. However, this is definitely far removed from my own expertise so there may be some parts of this which are not too good. I am following various algebra books in putting this together.

Consider the notion of splitting fields. It is desired to show that any two are isomorphic, meaning that there exists a one to one and onto mapping from one to the other which preserves all the algebraic structure. To begin with, here is a theorem about extending homomorphisms. [20]

Definition 10.4.1 Suppose \( F, \overline{F} \) are two fields and that \( f : F \rightarrow \overline{F} \) is a homomorphism. This means that
\[ f(xy) = f(x)f(y), \ f(x+y) = f(x) + f(y) \]
An isomorphism is a homomorphism which is one to one and onto. A monomorphism is a homomorphism which is one to one. An automorphism is an isomorphism of a single field. Sometimes people use the symbol \( \simeq \) to indicate something is an isomorphism. Then if \( p(x) \in F[x] \), say

\[
p(x) = \sum_{k=0}^{n} a_k x^k,
\]

\( \tilde{p}(x) \) will be the polynomial in \( \overline{F}[x] \) defined as

\[
\tilde{p}(x) = \sum_{k=0}^{n} f(a_k) x^k.
\]

Also consider \( f \) as a homomorphism of \( F[x] \) and \( \overline{F}[x] \) in the obvious way.

\[
f(p(x)) = \tilde{p}(x)
\]

It is clear that if \( f \) is an isomorphism of the two fields \( F, \overline{F} \), then it is also an isomorphism of the commutative rings \( F[x], \overline{F}[x] \) meaning that it is one to one and onto and preserves the two operations of addition and multiplication.

The following is a nice theorem which will be useful.

**Theorem 10.4.2** Let \( F \) be a field and let \( r \) be algebraic over \( F \). Let \( p(x) \) be the minimum polynomial of \( r \). Thus \( p(r) = 0 \) and \( p(x) \) is monic and no nonzero polynomial having coefficients in \( F \) of smaller degree has \( r \) as a root. In particular, \( p(x) \) is irreducible over \( F \). Then define \( f: F[x] \to F[r] \), the polynomials in \( r \) by

\[
f\left( \sum_{i=0}^{m} a_i x^i \right) \equiv \sum_{i=0}^{m} a_i r^i
\]

Then \( f \) is a homomorphism. Also, defining \( g: F[x] / (p(x)) \) by

\[
g([q(x)]) = f(q(x)) = q(r)
\]

it follows that \( g \) is an isomorphism from the field \( F[x] / (p(x)) \) to \( F[r] \).

**Proof:** First of all, consider why \( f \) is a homomorphism. The preservation of sums is obvious. Consider products.

\[
f\left( \sum_{i} a_i x^i \sum_{j} b_j x^j \right) = f\left( \sum_{i,j} a_i b_j x^{i+j} \right) = \sum_{i,j} a_i b_j r^{i+j}
\]

\[
= \sum_{i} a_i r^i \sum_{j} b_j r^j = f\left( \sum_{i} a_i x^i \right) f\left( \sum_{j} b_j x^j \right)
\]

Thus it is clear that \( f \) is a homomorphism.

First consider why \( g \) is even well defined. If \( [q(x)] = [q_1(x)] \), this means that

\[
q_1(x) - q(x) = p(x) l(x)
\]

for some \( l(x) \in F[x] \). Therefore,

\[
f(q_1(x)) = f(q(x)) + f(p(x)) l(x)
\]

\[
= f(q(x)) + f(p(x)) f(l(x))
\]

\[
= q(r) + p(r) l(r) = q(r) = f(q(x))
\]
Now from this, it is obvious that \( g \) is a homomorphism.

\[
g([q(x)][q_1(x)]) = g([q(x)q_1(x)]) = f(q(x)q_1(x)) = q(r)q_1(r)
\]

Similarly, \( g \) preserves sums. Now why is \( g \) one to one? It suffices to show that if \( g([q(x)]) = 0 \), then \([q(x)] = 0\). Suppose then that

\[
g([q(x)]) = q(r) = 0
\]

Then

\[
q(x) = p(x)l(x) + \rho(x)
\]

where the degree of \( \rho(x) \) is less than the degree of \( p(x) \) or else \( \rho(x) = 0 \). If \( \rho(x) \neq 0 \), then it follows that

\[
\rho(r) = 0
\]

and \( \rho(x) \) has smaller degree than that of \( p(x) \) which contradicts the definition of \( p(x) \) as the minimum polynomial of \( r \). Thus \( q(x) = p(x)l(x) \) and so \([q(x)] = 0\). Since \( p(x) \) is irreducible, \( \mathbb{F}[x]/(p(x)) \) is a field. It is clear that \( g \) is onto. Therefore, \( \mathbb{F}[r] \) is a field also. (This was shown earlier by different reasoning.) ■

Here is a diagram of what the following theorem says.

### Extending \( f \) to \( g \)

\[
\begin{array}{cccc}
\mathbb{F} & f & \mathbb{F} \\
p(x) \in \mathbb{F}[x] & \xrightarrow{f} & \mathbb{F}[x] \\
p(x) = \sum_{k=0}^{n} a_k x^k & \xrightarrow{f} & \sum_{k=0}^{n} f(a_k) x^k = \overline{p}(x) \\
p(r) = 0 & \xrightarrow{f} & \overline{p}(\overline{r}) = 0 \\
\mathbb{F}[r] & \xrightarrow{g} & \overline{\mathbb{F}}[\overline{r}]
\end{array}
\]

The idea illustrated is the following question: For \( r \) algebraic over \( \mathbb{F} \) and \( f \) an isomorphism of \( \mathbb{F} \) and \( \overline{\mathbb{F}} \), when does there exist \( \overline{r} \) algebraic over \( \overline{\mathbb{F}} \) and an isomorphism of \( \mathbb{F}[r] \) and \( \overline{\mathbb{F}}[\overline{r}] \) which extends \( f \)? This is the content of the following theorem.

**Theorem 10.4.3** Let \( f : \mathbb{F} \to \overline{\mathbb{F}} \) be an isomorphism and let \( r \) be algebraic over \( \mathbb{F} \) with minimal polynomial \( p(x) \). Then the following are equivalent.

1. There exists \( \overline{r} \) algebraic over \( \overline{\mathbb{F}} \) such that \( \overline{p}(\overline{r}) = 0 \) in which case \( \overline{p}(x) \) is the minimum polynomial of \( \overline{r} \).
2. There exists \( g : \mathbb{F}[r] \to \overline{\mathbb{F}}[\overline{r}] \) an isomorphism which extends \( f \) such that \( g(r) = \overline{r} \). In this case, there is only one such isomorphism.

**Proof:** 2.)\( \Rightarrow \)1.) Let \( g(r) = \overline{g}(\overline{r}) \) with \( g \) an isomorphism extending \( f \), \( g(r) = \overline{g}(\overline{r}) \). Then since it is an isomorphism,

\[
0 = g(p(r)) = \overline{p}(\overline{g}(\overline{r})) = \overline{p}(\overline{r}) \quad \text{(*)}
\]

Define \( \beta \) as \( \beta([k(x)]) = \overline{k}(\overline{r}) \) relative to this \( \overline{r} \) \( \equiv g(r) \) and let \( \alpha : \mathbb{F}[x]/(p(x)) \to \mathbb{F}[r] \) be the isomorphism mentioned in Theorem 10.3.4 called \( g \) there, given by \( \alpha([k(x)]) = k(r) \). Thus

\[
\mathbb{F}[r] \xrightarrow{\alpha} \mathbb{F}[x]/(p(x)) \xrightarrow{\beta} \overline{\mathbb{F}}[\overline{r}]
\]

Then if \( \beta \) is a well defined homomorphism, it follows that \( g \) must equal \( \beta \circ \alpha^{-1} \) because

\[
\beta \circ \alpha^{-1}(k(r)) \equiv \beta([k(x)]) \equiv \overline{k}(\overline{r}) \equiv \overline{g}(\overline{r}) = g(k(r))
\]
This is because $g$ is a homomorphism which takes $r$ to $\bar{r}$. It only remains to verify that $\beta$ is well defined.

Why is $\beta$ well defined? Suppose $[k(x)] = [k'(x)]$ so that $k(x) - k'(x) = l(x)p(x)$. Then since $f$ is a homomorphism, it follows from * that

$$
\bar{k}(x) - \bar{k}'(x) = \bar{l}(x)\bar{p}(\bar{x}) \Rightarrow \bar{k}(\bar{r}) - \bar{k}'(\bar{r}) = \bar{l}(\bar{r})\bar{p}(\bar{r}) = 0
$$

so $\beta$ is indeed well defined. It is clear from the definition that $\beta$ is a homomorphism.

1.) $\Rightarrow$ 2.) Next suppose there exists $\bar{r}$ algebraic over $F$ such that $\bar{p}(\bar{r}) = 0$. Why is $\bar{p}(x)$ the minimum polynomial of $\bar{r}$? Call it $\bar{q}(x)$. There is no loss of generality because $f$ is an isomorphism so the minimum polynomial can be written this way. Then $\beta([q(x)]) = \bar{q}(\bar{r}) = 0 = \bar{p}(\bar{r})$. Then $\bar{p}(x) = \bar{q}(x)\bar{m}(x) + R(x)$ where the degree of $R(x)$ is less than the degree of $\bar{q}(x)$ or equal to zero and so $R(\bar{r}) = 0$ which is contrary to $\bar{q}(x)$ being minimum polynomial for $\bar{r}$ unless $R(x) = 0$. Therefore, $R(x) = 0$. It follows that, since $f$ is an isomorphism, we have $p(x) = q(x)m(x)$ contrary to $p(x)$ being the minimum polynomial for $r$. Indeed, if the degree of $q(x)$ is less than that of $p(x)$, we have $0 = p(r) = q(r)m(r)$ and so one of $q(r)$, $m(r)$ equals 0 contrary to $p(x)$ having smallest possible degree for sending $r$ to 0. Thus the degree of $q(x)$ is the same as the degree of $p(x)$ and since both are monic by definition, $m(x) = 1$. Hence $p(x) = q(x)$ and so $\bar{p}(x) = \bar{q}(x)$.

Now let $\alpha, \beta$ be defined as above. It was shown above that $\beta$ is a well defined homomorphism. It is also clear that $\beta$ is onto. It only remains to verify that $\beta$ is one to one and when this is done, the isomorphism will be $\beta \circ \alpha^{-1}$. Suppose $\beta([k(x)]) = \bar{k}(\bar{r}) = 0$. Does it follow that $[k(x)] = 0$? By assumption, $\bar{p}(\bar{r}) = 0$ and also,

$$
\bar{k}(x) = \bar{p}(x)\bar{l}(x) + \bar{p}(x)
$$

(*)

where the degree of $\bar{p}(x)$ is less than the degree of $\bar{p}(x)$ which is the same as the degree of $p(x)$ or else it equals 0. It follows that $\bar{p}(\bar{r}) = 0$ and this is a contradiction because $\bar{p}(x)$ is the minimum polynomial for $\bar{r}$ which was shown above. Hence $\bar{k}(x) = \bar{p}(x)\bar{l}(x)$ and since $f$ is an isomorphism, this says that $k(x) = p(x)l(x)$ and so $[k(x)] = 0$. Hence $\beta$ is indeed one to one and so an example of $\eta$ would be $\beta \circ \alpha^{-1}$. Also $\beta \circ \alpha^{-1}(r) = \beta([x]) = \bar{r}$. $\blacksquare$

What is the meaning of the above in simple terms? It says that the monomorphisms from $F[r]$ to a field $K$ containing $F$ correspond to the roots of $p(x)$ in $K$. That is, for each root of $p(x)$, there is a monomorphism and for each monomorphism, there is a root. Also, for each root $\bar{r}$ of $\bar{p}(x)$ in $K$, there is an isomorphism from $F[r]$ to $F[\bar{r}]$. Here $p(x)$ is the minimum polynomial for $r$.

Note that if $p(x)$ is a monic irreducible polynomial, then it is the minimum polynomial for each of its roots. Consider why this is. If $r$ is a root of $p(x)$, then let $q(x)$ be the minimum polynomial for $r$. Then

$$p(x) = q(x)k(x) + R(x)$$

where $R(x)$ is 0 or else has smaller degree than $q(x)$. However, $R(r) = 0$ and this contradicts $q(x)$ being the minimum polynomial of $r$. Hence $q(x)$ divides $p(x)$ or else $k(x) = 1$. The latter possibility must be the case because $p(x)$ is irreducible.

This is the situation which is about to be considered. It involves the splitting fields $K, \mathbb{K}$ of $p(x), \bar{p}(x)$ where $\eta$ is an isomorphism of $F$ and $\bar{F}$ as described above. See [24]. Here is a little diagram which describes what this theorem says.

**Definition 10.4.4** The symbol $[K : F]$ where $K$ is a field extension of $F$ means the dimension of the vector space $K$ with field of scalars $F$.

$$
\begin{array}{c c c c c c c c c}
\mathbb{F} & \mathbb{F} & \mathbb{F} \\
\mathbb{F}[\bar{r}_1, \ldots, \bar{r}_n] & \mathbb{F}[r_1, \ldots, r_n] & \mathbb{F}[r_1, \ldots, r_n] \\
p(x) & \eta \simeq & \bar{p}(x) \\
\eta \circ \eta(p(x)) = \bar{p}(x) & \bar{p}(x) & \bar{p}(x) \\
m & m \leq [K : F] & m \leq [K : F], \bar{r}_i \neq \bar{r}_j
\end{array}
$$
Theorem 10.4.5 Let $\eta$ be an isomorphism from $\mathbb{F}$ to $\bar{\mathbb{F}}$ and let $\mathbb{K} = \mathbb{F}[r_1, \ldots, r_n], \bar{\mathbb{K}} = \bar{\mathbb{F}}[\bar{r}_1, \ldots, \bar{r}_n]$ be splitting fields of $p(x)$ and $\bar{p}(x)$ respectively. Then there exist at most $[\mathbb{K} : \mathbb{F}]$ isomorphisms $\zeta_i : \mathbb{K} \to \mathbb{K}$ which extend $\eta$. If $\{r_1, \ldots, r_n\}$ are distinct, then there exist exactly $[\mathbb{K} : \mathbb{F}]$ isomorphisms of the above sort. In either case, the two splitting fields are isomorphic with any of these $\zeta_i$ serving as an isomorphism.

Proof: Suppose $[\mathbb{K} : \mathbb{F}] = 1$. Say a basis for $\mathbb{K}$ is $\{r\}$. Then $\{1, r\}$ is dependent and so there exist $a, b \in \mathbb{F}$, not both zero such that $a + br = 0$. Then it follows that $r \in \mathbb{F}$ and so in this case $\mathbb{F} = \mathbb{K}$. Then the isomorphism which extends $\eta$ is just $\eta$ itself and there is exactly 1 isomorphism.

Next suppose $[\mathbb{K} : \mathbb{F}] > 1$. Then $p(x)$ has an irreducible factor over $\mathbb{F}$ of degree larger than 1, $q(x)$. If not, you would have

$$p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_n$$

and it would factor as

$$= (x - r_1) \cdots (x - r_n)$$

with each $r_j \in \mathbb{F}$, so $\mathbb{F} = \mathbb{K}$ contrary to $[\mathbb{K} : \mathbb{F}] > 1$. Without loss of generality, let the roots of $q(x)$ in $\mathbb{K}$ be $\{r_1, \ldots, r_m\}$. Thus

$$q(x) = \prod_{i=1}^{m} (x - r_i), \quad p(x) = \prod_{i=1}^{n} (x - r_i)$$

Now $\bar{q}(x)$ defined analogously to $\bar{p}(x)$, also has degree at least 2. Furthermore, it divides $\bar{p}(x)$ all of whose roots are in $\bar{\mathbb{K}}$. This is obvious because $\eta$ is an isomorphism. You have

$$l(x)q(x) = p(x) \quad \text{so} \quad \bar{l}(x)\bar{q}(x) = \bar{p}(x).$$

Denote the roots of $\bar{q}(x)$ in $\bar{\mathbb{K}}$ as $\{\bar{r}_1, \ldots, \bar{r}_m\}$ where they are counted according to multiplicity.

Then from Theorem 10.4.3, there exist $k \leq m$ one to one homomorphisms (monomorphisms) $\zeta_i$ mapping $\mathbb{F}[r_1]$ to $\mathbb{K} \equiv \mathbb{F} [\bar{r}_1, \ldots, \bar{r}_n]$, one for each distinct root of $\bar{q}(x)$ in $\bar{\mathbb{K}}$. If the roots of $\bar{p}(x)$ are distinct, then this is sufficient to imply that the roots of $\bar{q}(x)$ are also distinct, and $k = m$, the dimension of $q(x)$. Otherwise, maybe $k < m$. (It is conceivable that $\bar{q}(x)$ might have repeated roots in $\bar{\mathbb{K}}$.) Then

$$[\mathbb{K} : \mathbb{F}] = [\mathbb{K} : \mathbb{F}[r_1]] [\mathbb{F}[r_1] : \mathbb{F}]$$

and since the degree of $q(x) > 1$ and $q(x)$ is irreducible, this shows that $[\mathbb{F}[r_1] : \mathbb{F}] = m > 1$ and so

$$[\mathbb{K} : \mathbb{F}[r_1]] < [\mathbb{K} : \mathbb{F}]$$

Therefore, by induction, using Theorem 10.4.3, each of these $k \leq m = [\mathbb{F}[r_1] : \mathbb{F}]$ one to one homomorphisms extends to an isomorphism from $\mathbb{K}$ to $\mathbb{K}$ and for each of these $\zeta_i$, there are no more than $[\mathbb{K} : \mathbb{F}[r_1]]$ of these isomorphisms extending $\mathbb{F}$. If the roots of $\bar{p}(x)$ are distinct, then there are exactly $m$ of these $\zeta_i$ and for each, there are $[\mathbb{K} : \mathbb{F}[r_1]]$ extensions. Therefore, if the roots of $\bar{p}(x)$ are distinct, this has identified

$$[\mathbb{K} : \mathbb{F}[r_1]] m = [\mathbb{K} : \mathbb{F}[r_1]] [\mathbb{F}[r_1] : \mathbb{F}] = [\mathbb{K} : \mathbb{F}]$$

isomorphisms of $\mathbb{K}$ to $\bar{\mathbb{K}}$ which agree with $\eta$ on $\mathbb{F}$. If the roots of $\bar{p}(x)$ are not distinct, then maybe there are fewer than $[\mathbb{K} : \mathbb{F}]$ extensions of $\eta$.

Is this all of them? Suppose $\zeta$ is such an isomorphism of $\mathbb{K}$ and $\bar{\mathbb{K}}$. Then consider its restriction to $\mathbb{F}[r_1]$. By Theorem 10.4.3, this restriction must coincide with one of the $\zeta_i$ chosen earlier. Then by induction, $\zeta$ is one of the extensions of the $\zeta_i$ just mentioned. 

Definition 10.4.6 Let $\mathbb{K}$ be a finite dimensional extension of a field $\mathbb{F}$ such that every element of $\mathbb{K}$ is algebraic over $\mathbb{F}$, that is, each element of $\mathbb{K}$ is a root of some polynomial in $\mathbb{F}[x]$. Then $\mathbb{K}$ is called a normal extension if for every $k \in \mathbb{K}$ all roots of the minimum polynomial of $k$ are contained in $\mathbb{K}$. 
So what are some ways to tell that a field is a normal extension? It turns out that if \( \mathbb{K} \) is a splitting field of \( f(x) \in \mathbb{F}[x] \), then \( \mathbb{K} \) is a normal extension. I found this in [20]. This is an amazing result.

**Proposition 10.4.7** Let \( \mathbb{K} \) be a splitting field of \( f(x) \in \mathbb{F}[x] \). Then \( \mathbb{K} \) is a normal extension. In fact, if \( \mathbb{L} \) is an intermediate field between \( \mathbb{F} \) and \( \mathbb{K} \), then \( \mathbb{L} \) is also a normal extension of \( \mathbb{F} \).

**Proof:** Let \( r \in \mathbb{K} \) be a root of \( g(x) \), an irreducible monic polynomial in \( \mathbb{F}[x] \). It is required to show that every other root of \( g(x) \) is in \( \mathbb{K} \). Let the roots of \( g(x) \) in a splitting field be \( \{r_1 = r, r_2, \ldots, r_m\} \). Now \( g(x) \) is the minimum polynomial of \( r_j \) over \( \mathbb{F} \) because \( g(x) \) is irreducible. Recall why this was. If \( p(x) \) is the minimum polynomial of \( r_j \),

\[
g(x) = p(x) l(x) + r(x)
\]

where \( r(x) \) either is 0 or it has degree less than the degree of \( p(x) \). However, \( r(r_j) = 0 \) and this is impossible if \( p(x) \) is the minimum polynomial. Hence \( r(x) = 0 \) and now it follows that \( g(x) \) was not irreducible unless \( l(x) = 1 \).

By Theorem [10.3.2], there exists an isomorphism \( \eta \) of \( \mathbb{F}[r_1] \) and \( \mathbb{F}[r_j] \) which fixes \( \mathbb{F} \) and maps \( r_1 \) to \( r_j \). Now \( \mathbb{K}[r_1] \) and \( \mathbb{K}[r_j] \) are splitting fields of \( f(x) \) over \( \mathbb{F}[r_1] \) and \( \mathbb{F}[r_j] \) respectively. By Theorem [10.4.1], the two fields \( \mathbb{K}[r_1] \) and \( \mathbb{K}[r_j] \) are isomorphic, the isomorphism, \( \zeta \) extending \( \eta \).

Hence

\[
[\mathbb{K}[r_1] : \mathbb{K}] = [\mathbb{K}[r_j] : \mathbb{K}]
\]

But \( r_1 \in \mathbb{K} \) and so \( \mathbb{K}[r_1] = \mathbb{K} \). Therefore, \( \mathbb{K} = \mathbb{K}[r_j] \) and so \( r_j \) is also in \( \mathbb{K} \). Thus all the roots of \( g(x) \) are actually in \( \mathbb{K} \). Consider the last assertion.

Suppose \( r = r_1 \in \mathbb{L} \) where the minimum polynomial for \( r \) is denoted by \( q(x) \). Then letting the roots of \( q(x) \) in \( \mathbb{K} \) be \( \{r_1, \ldots, r_m\} \). By Theorem [10.4.2] applied to the identity map on \( \mathbb{L} \), there exists an isomorphism \( \theta : \mathbb{L}[r_1] \to \mathbb{L}[r_j] \) which fixes \( \mathbb{L} \) and takes \( r_1 \) to \( r_j \). But this implies that

\[
1 = [\mathbb{L}[r_1] : \mathbb{L}] = [\mathbb{L}[r_j] : \mathbb{L}]
\]

Hence \( r_j \in \mathbb{L} \) also. Since \( r \) was an arbitrary element of \( \mathbb{L} \), this shows that \( \mathbb{L} \) is normal. \( \blacksquare \)

**Definition 10.4.8** When you have \( \mathbb{F}[a_1, \ldots, a_m] \) with each \( a_i \) algebraic so that \( \mathbb{F}[a_1, \ldots, a_m] \) is a field, you could consider

\[
f(x) \equiv \prod_{i=1}^{m} f_i(x)
\]

where \( f_i(x) \) is the minimum polynomial of \( a_i \). Then if \( \mathbb{K} \) is a splitting field for \( f(x) \), this \( \mathbb{K} \) is called the normal closure. It is at least as large as \( \mathbb{F}[a_1, \ldots, a_m] \) and it has the advantage of being a normal extension.
Part II

Analysis And Geometry In Linear Algebra
Chapter 11

Normed Linear Spaces

In addition to the algebraic aspects of linear algebra presented earlier, there are many analytical and geometrical concepts which are usually included. This material involves the special fields \( \mathbb{R} \) and \( \mathbb{C} \) instead of general fields. It is these things which are typically generalized in functional analysis. The main new idea is that the notion of distance is included. This allows one to consider continuity, compactness, and many other topics from calculus. First is a general treatment of the notion of distance which has nothing to do with linear algebra but is a useful part of the vocabulary leading most efficiently to the inclusion of analytical topics.

11.1 Metric Spaces

This section is here to provide definitions and main theorems about fundamental analytical ideas and terminology.

11.1.1 Open And Closed Sets, Sequences, Limit Points, Completeness

It is most efficient to discuss things in terms of abstract metric spaces to begin with.

**Definition 11.1.1** A non empty set \( X \) is called a metric space if there is a function \( d : X \times X \to [0, \infty) \) which satisfies the following axioms.

1. \( d(x, y) = d(y, x) \)
2. \( d(x, y) \geq 0 \) and equals 0 if and only if \( x = y \)
3. \( d(x, y) + d(y, z) \geq d(x, z) \)

This function \( d \) is called the metric. We often refer to it as the distance also.

**Definition 11.1.2** An open ball, denoted as \( B(x, r) \) is defined as follows.

\[
B(x, r) \equiv \{ y : d(x, y) < r \}
\]

A set \( U \) is said to be open if whenever \( x \in U \), it follows that there is \( r > 0 \) such that \( B(x, r) \subseteq U \). More generally, a point \( x \) is said to be an interior point of \( U \) if there exists such a ball. In words, an open set is one for which every point is an interior point.

For example, you could have \( X \) be a subset of \( \mathbb{R} \) and \( d(x, y) = |x - y| \).

Then the first thing to show is the following.

**Proposition 11.1.3** An open ball is an open set.
Proof: Suppose \( y \in B(x, r) \). We need to verify that \( y \) is an interior point of \( B(x, r) \). Let \( \delta = r - d(x, y) \). Then if \( z \in B(y, \delta) \), it follows that

\[
d(z, x) \leq d(z, y) + d(y, x) < \delta + d(y, x) = r - d(x, y) + d(y, x) = r
\]

Thus \( y \in B(y, \delta) \subseteq B(x, r) \). ■

Definition 11.1.4 Let \( S \) be a nonempty subset of a metric space. Then \( p \) is a limit point (accumulation point) of \( S \) if for every \( r > 0 \) there exists a point different than \( p \) in \( B(p, r) \cap S \). Sometimes people denote the set of limit points as \( S' \).

A related idea is the notion of the limit of a sequence. Recall that a sequence is really just a mapping from \( \mathbb{N} \) to \( X \). We write them as \( \{x_n\} \) or \( \{x_n\}_{n=1}^{\infty} \) if we want to emphasize the values of \( n \). Then the following definition is what it means for a sequence to converge.

Definition 11.1.5 We say that \( x = \lim_{n \to \infty} x_n \) when for every \( \varepsilon > 0 \) there exists \( N \) such that if \( n \geq N \), then

\[
d(x, x_n) < \varepsilon
\]

Often we write \( x_n \to x \) for short. This is equivalent to saying

\[
\lim_{n \to \infty} d(x, x_n) = 0.
\]

Proposition 11.1.6 The limit is well defined. That is, if \( x, x' \) are both limits of a sequence, then \( x = x' \).

Proof: From the definition, there exist \( N, N' \) such that if \( n \geq N \), then \( d(x, x_n) < \varepsilon/2 \) and if \( n \geq N' \), then \( d(x, x_n) < \varepsilon/2 \). Then let \( M \geq \max(N, N') \). Let \( n > M \). Then

\[
d(x, x') \leq d(x, x_n) + d(x_n, x') < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon
\]

Since \( \varepsilon \) is arbitrary, this shows that \( x = x' \) because \( d(x, x') = 0 \). ■

Next there is an important theorem about limit points and convergent sequences.

Theorem 11.1.7 Let \( S \neq \emptyset \). Then \( p \) is a limit point of \( S \) if and only if there exists a sequence of distinct points of \( S \), \( \{x_n\} \) none of which equal \( p \) such that \( \lim_{n \to \infty} x_n = p \).

Proof: \( \implies \) Suppose \( p \) is a limit point. Why does there exist the promised convergent sequence? Let \( x_1 \in B(p, 1) \cap S \) such that \( x_1 \neq p \). If \( x_1, \ldots, x_n \) have been chosen, let \( x_{n+1} \neq p \) be in \( B(p, \delta_{n+1}) \cap S \) where \( \delta_{n+1} = \min \left\{ \frac{1}{n+1}, d(x_i, p), i = 1, 2, \ldots, n \right\} \). Then this constructs the necessary convergent sequence.

\( \implies \) Conversely, if such a sequence \( \{x_n\} \) exists, then for every \( r > 0 \), \( B(p, r) \) contains \( x_n \in S \) for all \( n \) large enough. Hence, \( p \) is a limit point because none of these \( x_n \) are equal to \( p \). ■

Definition 11.1.8 A set \( H \) is closed means \( H^C \) is open.

Note that this says that the complement of an open set is closed. If \( V \) is open, then the complement of its complement is itself. Thus \( (V^C)^C = V \) an open set. Hence \( V^C \) is closed.

Then the following theorem gives the relationship between closed sets and limit points.

Theorem 11.1.9 A set \( H \) is closed if and only if it contains all of its limit points.

Proof: \( \implies \) Let \( H \) be closed and let \( p \) be a limit point. We need to verify that \( p \in H \). If it is not, then since \( H \) is closed, its complement is open and so there exists \( \delta > 0 \) such that \( B(p, \delta) \cap H = \emptyset \). However, this prevents \( p \) from being a limit point.

\( \implies \) Next suppose \( H \) has all of its limit points. Why is \( H^C \) open? If \( p \in H^C \) then it is not a limit point and so there exists \( \delta > 0 \) such that \( B(p, \delta) \) has no points of \( H \). In other words, \( H^C \) is open. Hence \( H \) is closed. ■
11.1. METRIC SPACES

Corollary 11.1.10 A set $H$ is closed if and only if whenever $\{h_n\}$ is a sequence of points of $H$ which converges to a point $x$, it follows that $x \in H$.

Proof: $\implies$ Suppose $H$ is closed and $h_n \to x$. If $x \in H$ there is nothing left to show. If $x \notin H$, then from the definition of limit, it is a limit point of $H$. Hence $x \in H$ after all.

$\impliedby$ Suppose the limit condition holds, why is $H$ closed? Let $x \notin H$ the set of limit points of $H$. By Theorem 11.1.7 there exists a sequence of points of $H$, $\{h_n\}$ such that $h_n \to x$. Then by assumption, $x \in H$. Thus $H$ contains all of its limit points and so it is closed by Theorem 11.1.9. ■

Next is the important concept of a subsequence.

Definition 11.1.11 Let $\{x_n\}_{n=1}^{\infty}$ be a sequence. Then if $n_1 < n_2 < \cdots$ is a strictly increasing sequence of indices, we say $\{x_{n_k}\}_{k=1}^{\infty}$ is a subsequence of $\{x_n\}_{n=1}^{\infty}$.

The really important thing about subsequences is that they preserve convergence.

Theorem 11.1.12 Let $\{x_{n_k}\}$ be a subsequence of a convergent sequence $\{x_n\}$ where $x_n \to x$. Then

$$\lim_{k \to \infty} x_{n_k} = x$$

also.

Proof: Let $\varepsilon > 0$ be given. Then there exists $N$ such that

$$d(x_n, x) < \varepsilon \text{ if } n \geq N.$$ 

It follows that if $k \geq N$, then $n_k \geq N$ and so

$$d(x_{n_k}, x) < \varepsilon \text{ if } k \geq N.$$ 

This is what it means to say $\lim_{k \to \infty} x_{n_k} = x$. ■

Another useful idea is the distance to a set.

Definition 11.1.13 Let $(X, d)$ be a metric space and let $S$ be a nonempty set in $X$. Then

$$\text{dist}(x, S) \equiv \inf \{d(x, y) : y \in S\}.$$ 

The following lemma is the fundamental result.

Lemma 11.1.14 The function, $x \to \text{dist}(x, S)$ is continuous and in fact satisfies

$$|\text{dist}(x, S) - \text{dist}(y, S)| \leq d(x, y).$$ 

Proof: Suppose $\text{dist}(x, S)$ is as least as large as $\text{dist}(y, S)$. Then pick $z \in S$ such that $d(y, z) \leq \text{dist}(y, S) + \varepsilon$. Then

$$|\text{dist}(x, S) - \text{dist}(y, S)| = \text{dist}(x, S) - \text{dist}(y, S) \leq d(x, z) - (d(y, z) - \varepsilon) = d(x, z) - d(y, z) + \varepsilon \leq d(x, y) + d(y, z) - d(y, z) + \varepsilon = d(x, y) + \varepsilon.$$ 

Since $\varepsilon > 0$ is arbitrary, this proves the lemma. The argument is the same if $\text{dist}(x, S) \leq \text{dist}(y, S)$. Just switch the roles of $x$ and $y$. ■
11.1.2 Cauchy Sequences, Completeness

Of course it does not go the other way. For example, you could let \( x_n = (-1)^n \) and it has a convergent subsequence but fails to converge. Here \( d(x, y) = |x - y| \) and the metric space is just \( \mathbb{R} \).

However, there is a kind of sequence for which it does go the other way. This is called a Cauchy sequence.

**Definition 11.1.15** \( \{x_n\} \) is called a Cauchy sequence if for every \( \varepsilon > 0 \) there exists \( N \) such that if \( m, n \geq N \), then

\[
 d(x_n, x_m) < \varepsilon
\]

Now the major theorem about this is the following.

**Theorem 11.1.16** Let \( \{x_n\} \) be a Cauchy sequence. Then it converges if and only if any subsequence converges.

**Proof:** \( \implies \) This was just done above. \( \iff \) Suppose now that \( \{x_n\} \) is a Cauchy sequence and \( \lim_{k \to \infty} x_{n_k} = x \). Then there exists \( N_1 \) such that if \( k > N_1 \), then \( d(x_{n_k}, x) < \varepsilon/2 \). From the definition of what it means to be Cauchy, there exists \( N_2 \) such that if \( m, n \geq N_2 \), then \( d(x_m, x_n) < \varepsilon/2 \). Let \( N = \max(N_1, N_2) \). Then if \( k \geq N \), then \( n_k \geq N \) and so

\[
 d(x, x_k) \leq d(x, x_{n_k}) + d(x_{n_k}, x_k) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon
\]

It follows from the definition that \( \lim_{k \to \infty} x_k = x. \) ■

**Definition 11.1.17** A metric space is said to be complete if every Cauchy sequence converges.

Another nice thing to note is this.

**Proposition 11.1.18** If \( \{x_n\} \) is a sequence and if \( p \) is a limit point of the set \( S = \bigcup_{n=1}^{\infty} \{x_n\} \) then there is a subsequence \( \{x_{n_k}\} \) such that \( \lim_{k \to \infty} x_{n_k} = x \).

**Proof:** By Theorem 11.1.14, there exists a sequence of distinct points of \( S \) denoted as \( \{y_k\} \) such that none of them equal \( p \) and \( \lim_{k \to \infty} y_k = p \). Thus \( B(p, r) \) contains infinitely many different points of the set \( D \), this for every \( r \). Let \( x_{n_1} \in B(p, 1) \) where \( n_1 \) is the first index such that \( x_{n_1} \in B(p, 1) \). Suppose \( x_{n_1}, \ldots, x_{n_k} \) have been chosen, the \( n_i \) increasing and let \( 1 > \delta_1 > \delta_2 > \cdots > \delta_k \) where \( x_{n_i} \in B(p, \delta_i) \). Then let

\[
 \delta_{k+1} \leq \min \left\{ \frac{1}{2^{k+1}}, d(p, x_{n_1}), \delta_j, j = 1, 2, \ldots, k \right\}
\]

Let \( x_{n_{k+1}} \in B(p, \delta_{k+1}) \) where \( n_{k+1} \) is the first index such that \( x_{n_{k+1}} \) is contained \( B(p, \delta_{k+1}) \). Then

\[
 \lim_{k \to \infty} x_{n_k} = p. \) ■

Another useful result is the following.

**Lemma 11.1.19** Suppose \( x_n \to x \) and \( y_n \to y \). Then \( d(x_n, y_n) \to d(x, y) \).

**Proof:** Consider the following.

\[
 d(x, y) \leq d(x, x_n) + d(x_n, y) \leq d(x, x_n) + d(x_n, y_n) + d(y_n, y)
\]

so

\[
 d(x, y) - d(x_n, y_n) \leq d(x, x_n) + d(y_n, y)
\]

Similarly

\[
 d(x_n, y_n) - d(x, y) \leq d(x, x_n) + d(y_n, y)
\]
and so
\[ |d(x_n, y_n) - d(x, y)| \leq d(x, x_n) + d(y_n, y) \]
and the right side converges to 0 as \( n \to \infty \).

First are some simple lemmas featuring one dimensional considerations. In these, the metric space is \( \mathbb{R} \) and the distance is given by
\[ d(x, y) = |x - y| \]
First recall the nested interval lemma. You should have seen something like it in calculus, but this is often not the case because there is much more interest in trivialities like integration techniques.

**Lemma 11.1.20** Let \([a_k, b_k] \supseteq [a_{k+1}, b_{k+1}]\) for all \( k = 1, 2, 3, \cdots \). Then there exists a point \( p \) in \( \bigcap_{k=1}^\infty [a_k, b_k] \).

**Proof:** We note that for any \( k, l, a_k \leq b_l \). Here is why. If \( k \leq l \), then
\[ a_k \leq b_k \leq b_l \]
If \( k > l \), then
\[ b_l \geq b_k \geq a_k \]
It follows that for each \( l \),
\[ \sup_k a_k \leq b_l \]
Hence \( \sup_k a_k \) is a lower bound to the set of all \( b_l \) and so it is no larger than the greatest lower bound. It follows that
\[ \sup_k a_k \leq \inf_l b_l \]
Pick \( x \in [\sup_k a_k, \inf_l b_l] \). Then for every \( k, a_k \leq x \leq b_k \). Hence \( x \in \bigcap_{k=1}^\infty [a_k, b_k] \).

**Lemma 11.1.21** The closed interval \([a, b]\) is compact. This means that if there is a collection of open intervals of the form \((a, b)\) whose union includes all of \([a, b]\), then in fact \([a, b]\) is contained in the union of finitely many of these open intervals.

**Proof:** Let \( C \) be a set of open intervals the union of which includes all of \([a, b]\) and suppose \([a, b]\) fails to admit a finite subcover. That is, no finite subset of \( C \) has union which contains \([a, b]\). Then this must be the case for one of the two intervals \([a, \frac{a+b}{2}]\) and \([\frac{a+b}{2}, b]\). Let \( I_1 \) be the one for which this is so. Then split it into two equal pieces like what was just done and let \( I_2 \) be a half for which there is no finite subcover of sets of \( C \). Continue this way. This yields a nested sequence of closed intervals \( I_1 \supseteq I_2 \supseteq \cdots \) and by the above lemma, there exists a point \( x \) in all of these intervals. There exists \( U \in C \) such that \( x \in U \). Thus
\[ x \in (a, b) \in C \]
However, for all \( n \) large enough, the length of \( I_n \) is less than \( \min(\{|x-a|, |x-b|\}) \). Hence \( I_n \) is actually contained in \((a, b) \in C \) contrary to the construction. Hence \([a, b]\) is compact after all.

As a useful corollary, this shows that \( \mathbb{R} \) is complete.

**Corollary 11.1.22** The real line \( \mathbb{R} \) is complete.

**Proof:** Suppose \( \{x_k\} \) is a Cauchy sequence in \( \mathbb{R} \). Then there exists \( M \) such that \( \{x_k\}_{k=1}^\infty \subseteq [-M, M] \). Why? If there is no convergent subsequence, then for each \( x \in [-M, M] \), there is an open set \((x - \delta_x, x + \delta_x)\) which contains \( x_k \) for only finitely many values of \( k \). Since \([-M, M]\) is compact, there are finitely many of these open sets whose union includes \([-M, M]\). This is a contradiction because \([-M, M]\) contains \( x_k \) for all \( k \in \mathbb{N} \) so at least one of the open sets must contain \( x_k \) for infinitely many \( k \). Thus there is a convergent subsequence. By Theorem the original Cauchy sequence converges to some \( x \in [-M, M] \).
Example 11.1.23 Let $n \in \mathbb{N}$. $\mathbb{C}^n$ with distance given by
\[
d(x, y) \equiv \max_{j \in \{1, \ldots, n\}} |x_j - y_j|
\]
is a complete space. Recall that $|a + jb| = \sqrt{a^2 + b^2}$. Then $\mathbb{C}^n$ is complete. Similarly $\mathbb{R}^n$ is complete.

To see that this is complete, let \(\{x^k\}_{k=1}^{\infty}\) be a Cauchy sequence. Observe that for each $j$, \(\{x^k_j\}_{k=1}^{\infty}\). That is, each component is a Cauchy sequence in $\mathbb{C}$. Next,
\[
\left| \text{Re} x^k_j - \text{Re} x^k_{j+p} \right| \leq \left| x^k_j - x^k_{j+p} \right|
\]
Therefore, \(\{\text{Re} x^k_j\}_{k=1}^{\infty}\) is a Cauchy sequence. Similarly \(\{\text{Im} x^k_j\}_{k=1}^{\infty}\) is a Cauchy sequence. It follows from completeness of $\mathbb{R}$ shown above, that these converge. Thus there exists $a_j, b_j$ such that
\[
\lim_{k \to \infty} \text{Re} x^k_j + i \text{Im} x^k_j = a_j + ib_j \equiv x
\]
and so $x^k \to x$ showing that $\mathbb{C}^n$ is complete. The same argument shows that $\mathbb{R}^n$ is complete. It is easier because you don’t need to fuss with real and imaginary parts.

11.1.3 Closure Of A Set

Next is the topic of the closure of a set.

Definition 11.1.24 Let $A$ be a nonempty subset of $(X, d)$ a metric space. Then $\overline{A}$ is defined to be the intersection of all closed sets which contain $A$. Note the whole space, $X$ is one such closed set which contains $A$. The whole space $X$ is closed because its complement is open, its complement being $\emptyset$. It is certainly true that every point of the empty set is an interior point because there are no points of $\emptyset$.

Lemma 11.1.25 Let $A$ be a nonempty set in $(X, d)$. Then $\overline{A}$ is a closed set and
\[
\overline{A} = A \cup A'
\]
where $A'$ denotes the set of limit points of $A$.

Proof: First of all, denote by $\mathcal{C}$ the set of closed sets which contain $A$. Then
\[
\overline{A} = \cap \mathcal{C}
\]
and this will be closed if its complement is open. However,
\[
\overline{A}^c = \cup \{H^c : H \in \mathcal{C}\}
\]
Each $H^c$ is open and so the union of all these open sets must also be open. This is because if $x$ is in this union, then it is in at least one of them. Hence it is an interior point of that one. But this implies it is an interior point of the union of them all which is an even larger set. Thus $\overline{A}$ is closed.

The interesting part is the next claim. First note that from the definition, $A \subseteq \overline{A}$ so if $x \in A$, then $x \in \overline{A}$. Now consider $y \in A'$ but $y \notin A$. If $y \notin \overline{A}$, a closed set, then there exists $B(y, r) \subseteq \overline{A}^c$. Thus $y$ cannot be a limit point of $A$, a contradiction. Therefore,
\[
A \cup A' \subseteq \overline{A}
\]
Next suppose $x \in \overline{A}$ and suppose $x \notin A$. Then if $B(x, r)$ contains no points of $A$ different than $x$, since $x$ itself is not in $A$, it would follow that $B(x, r) \cap A = \emptyset$ and so recalling that open balls are open, $B(x, r)^c$ is a closed set containing $A$ so from the definition, it also contains $\overline{A}$ which is contrary to the assertion that $x \notin \overline{A}$. Hence if $x \notin A$, then $x \in A'$ and so
\[
A \cup A' \supseteq \overline{A}
\]
11.1.4 Continuous Functions

The following is a fairly general definition of what it means for a function to be continuous. It includes everything seen in typical calculus classes as a special case.

**Definition 11.1.26** Let \( f : X \to Y \) be a function where \((X, d)\) and \((Y, \rho)\) are metric spaces. Then \( f \) is continuous at \( x \in X \) if and only if the following condition holds. For every \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that if \( d(\hat{x}, x) < \delta \), then \( \rho(f(\hat{x}), f(x)) < \varepsilon \). If \( f \) is continuous at every \( x \in X \) we say that \( f \) is continuous on \( X \).

For example, you could have a real valued function \( f(x) \) defined on an interval \([0, 1]\). In this case you would have \( X = [0, 1] \) and \( Y = \mathbb{R} \) with the distance given by \( d(x, y) = |x - y| \). Then the following theorem is the main result.

**Theorem 11.1.27** Let \( f : X \to Y \) where \((X, d)\) and \((Y, \rho)\) are metric spaces. Then the following are equivalent.

a. \( f \) is continuous at \( x \).

b. Whenever \( x_n \to x \), it follows that \( f(x_n) \to f(x) \).

Also, the following are equivalent.

c. \( f \) is continuous on \( X \).

d. Whenever \( V \) is open in \( Y \), it follows that \( f^{-1}(V) \equiv \{x : f(x) \in V\} \) is open in \( X \).

e. Whenever \( H \) is closed in \( Y \), it follows that \( f^{-1}(H) \) is closed in \( X \).

**Proof:**
a \( \implies \) b: Let \( f \) be continuous at \( x \) and suppose \( x_n \to x \). Then let \( \varepsilon > 0 \) be given. By continuity, there exists \( \delta > 0 \) such that if \( d(\hat{x}, x) < \delta \), then \( \rho(f(\hat{x}), f(x)) < \varepsilon \). Since \( x_n \to x \), it follows that there exists \( N \) such that if \( n \geq N \), then \( d(x_n, x) < \delta \) and so, if \( n \geq N \), it follows that \( \rho(f(x_n), f(x)) < \varepsilon \). Since \( \varepsilon > 0 \) is arbitrary, it follows that \( f(x_n) \to f(x) \).

b \( \implies \) a: Suppose b holds but \( f \) fails to be continuous at \( x \). Then there exists \( \varepsilon > 0 \) such that for all \( \delta > 0 \), there exists \( \hat{x} \) such that \( d(\hat{x}, x) < \delta \) but \( \rho(f(\hat{x}), f(x)) \geq \varepsilon \). Letting \( \delta = 1/n \), there exists \( x_n \) such that \( d(x_n, x) < 1/n \) but \( \rho(f(x_n), f(x)) \geq \varepsilon \). Now this is a contradiction because by assumption, the fact that \( x_n \to x \) implies that \( f(x_n) \to f(x) \). In particular, for large enough \( n \), \( \rho(f(x_n), f(x)) < \varepsilon \) contrary to the construction.

c \( \implies \) d: Let \( V \) be open in \( Y \). Let \( x \in f^{-1}(V) \) so that \( f(x) \in V \). Since \( V \) is open, there exists \( \varepsilon > 0 \) such that \( B(f(x), \varepsilon) \subseteq V \). Since \( f \) is continuous at \( x \), it follows that there exists \( \delta > 0 \) such that if \( \hat{x} \in B(x, \delta) \), then \( f(\hat{x}) \in B(f(x), \varepsilon) \subseteq V \). In other words, \( B(x, \delta) \subseteq f^{-1}(B(f(x), \varepsilon)) \subseteq f^{-1}(V) \) which shows that, since \( x \) was an arbitrary point of \( f^{-1}(V) \), every point of \( f^{-1}(V) \) is an interior point which implies \( f^{-1}(V) \) is open.

d \( \implies \) c: Let \( H \) be closed in \( Y \). Then \( f^{-1}(H)^C = f^{-1}(H^C) \) which is open by assumption. Hence \( f^{-1}(H)^C \) is closed because its complement is open.

e \( \implies \) d: Let \( V \) be open in \( Y \). Then \( f^{-1}(V)^C = f^{-1}(V^C) \) which is assumed to be closed. This is because the complement of an open set is a closed set.

d \( \implies \) c: Let \( x \in X \) be arbitrary. Is it the case that \( f \) is continuous at \( x \)? Let \( \varepsilon > 0 \) be given. Then \( B(f(x), \varepsilon) \) is an open set in \( Y \) and so \( x \in f^{-1}(B(f(x), \varepsilon)) \) which is given to be open. Hence there exists \( \delta > 0 \) such that \( x \in B(x, \delta) \subseteq f^{-1}(B(f(x), \varepsilon)) \). Thus, \( f(B(x, \delta)) \subseteq f(B(f(x), \varepsilon)) \) so \( \rho(f(\hat{x}), f(x)) < \varepsilon \). Thus \( f \) is continuous at \( x \) for every \( x \).

11.1.5 Separable Metric Spaces

**Definition 11.1.28** A metric space is called separable if there exists a countable dense subset \( D \). This means two things. First, \( D \) is countable, and second that if \( x \) is any point and \( r > 0 \), then \( B(x, r) \cap D \neq \emptyset \). A metric space is called completely separable if there exists a countable collection of nonempty open sets \( \mathcal{B} \) such that every open set is the union of some subset of \( \mathcal{B} \). This collection of open sets is called a countable basis.
A metric space is separable if and only if it is completely separable.

**Theorem 11.1.29** A metric space is separable if and only if it is completely separable.

**Proof:** $\iff$ Let $\mathcal{B}$ be the special countable collection of open sets and for each $B \in \mathcal{B}$, let $p_B$ be a point of $B$. Then let $\mathcal{P} = \{p_B : B \in \mathcal{B}\}$. If $B(x, r)$ is any ball, then it is the union of sets of $\mathcal{B}$ and so there is a point of $\mathcal{P}$ in it. Since $\mathcal{B}$ is countable, so is $\mathcal{P}$.

$\implies$ Let $D$ be the countable dense set and let $\mathcal{B} = \{B(d, r) : d \in D, r \in \mathbb{Q} \cap [0, \infty)\}$. Then $\mathcal{B}$ is countable because the Cartesian product of countable sets is countable. It suffices to show that every ball is the union of these sets. Let $B(x, R)$ be a ball. Let $y \in B(y, \delta) \subseteq B(x, R)$. Then there exists $d \in B \left(y, \frac{\delta}{10}\right)$. Let $\varepsilon \in \mathbb{Q}$ and $\frac{\delta}{10} < \varepsilon < \frac{5}{10}$. Then $y \in B(d, \varepsilon) \in \mathcal{B}$. Is $B(d, \varepsilon) \subseteq B(x, R)$? If so, then the desired result follows because this would show that every $y \in B(x, R)$ is contained in one of these sets of $\mathcal{B}$ which is contained in $B(x, R)$ showing that $B(x, R)$ is the union of sets of $\mathcal{B}$. Let $z \in B(d, \varepsilon) \subseteq B \left(d, \frac{\delta}{5}\right)$. Then

$$d(y, z) \leq d(y, d) + d(d, z) < \frac{\delta}{10} + \varepsilon < \frac{\delta}{10} + \frac{\delta}{5} < \delta$$

Hence $B(d, \varepsilon) \subseteq B(y, \delta) \subseteq B(x, r)$. Therefore, every ball is the union of sets of $\mathcal{B}$ and, since every open set is the union of balls, it follows that every open set is the union of sets of $\mathcal{B}$. $\blacksquare$

**Definition 11.1.30** Let $S$ be a nonempty set. Then a set of open sets $\mathcal{C}$ is called an open cover of $S$ if $\cup \mathcal{C} \supseteq S$. (It covers up the set $S$. Think lilly pads covering the surface of a pond.)

One of the important properties possessed by separable metric spaces is the Lindeloff property.

**Definition 11.1.31** A metric space has the Lindeloff property if whenever $\mathcal{C}$ is an open cover of a set $S$, there exists a countable subset of $\mathcal{C}$ denoted here by $\mathcal{B}$ such that $\mathcal{B}$ is also an open cover of $S$.

**Theorem 11.1.32** Every separable metric space has the Lindeloff property.

**Proof:** Let $\mathcal{C}$ be an open cover of a set $S$. Let $\mathcal{B}$ be a countable basis. Such exists by Theorem 11.1.29. Let $\mathcal{B}$ denote those sets of $\mathcal{B}$ which are contained in some set of $\mathcal{C}$. Thus $\mathcal{B}$ is a countable open cover of $S$. Now for $B \in \mathcal{B}$, let $U_B$ be a set of $\mathcal{C}$ which contains $B$. Letting $\mathcal{C}$ denote these sets $U_B$ it follows that $\mathcal{C}$ is countable and is an open cover of $S$. $\blacksquare$

**Definition 11.1.33** A Polish space is a complete separable metric space. These things turn out to be very useful in probability theory and in other areas.

### 11.1.6 Compact Sets

As usual, we are not worrying about empty sets.

**Definition 11.1.34** A metric space $K$ is compact if whenever $\mathcal{C}$ is an open cover of

$$K, (\cup \mathcal{C} \supseteq K, \text{ each set of } \mathcal{C} \text{ is open})$$

there exists a finite subset of $\mathcal{C} \{U_1, \cdots, U_n\}$ such that $K \subseteq \bigcup_{k=1}^{n} U_k$. In words, every open cover admits a finite sub-cover.

The above definition is equivalent to the same statement with the provision that each open set in $\mathcal{C}$ is an open ball. See Problem 23 on Page 273.

This is the real definition given above. However, in metric spaces, it is equivalent to another definition called sequentially compact.

**Definition 11.1.35** A metric space $K$ is sequentially compact means that whenever $\{x_n\} \subseteq K$, there exists a subsequence $\{x_{n_k}\}$ such that $\lim_{k \to \infty} x_{n_k} = x \in K$ for some point $x$. In words, every sequence has a subsequence which converges to a point in the set.
Definition 11.1.36  Let $X$ be a metric space. Then a finite set of points $\{x_1, \ldots, x_n\}$ is called an $\varepsilon$ net if

$$X \subseteq \bigcup_{k=1}^{n} B(x_k, \varepsilon)$$

If, for every $\varepsilon > 0$ a metric space has an $\varepsilon$ net, then we say that the metric space is totally bounded.

Lemma 11.1.37  If a metric space $(K, d)$ is sequentially compact, then it is separable and totally bounded.

Proof: Pick $x_1 \in K$. If $B(x_1, \varepsilon) \supseteq K$, then stop. Otherwise, pick $x_2 \notin B(x_1, \varepsilon)$. Continue this way. If $\{x_1, \ldots, x_n\}$ have been chosen, either $K \subseteq \bigcup_{k=1}^{n} B(x_k, \varepsilon)$ in which case, you have found an $\varepsilon$ net or this does not happen in which case, you can pick $x_{n+1} \notin \bigcup_{k=1}^{n} B(x_k, \varepsilon)$. The process must terminate since otherwise, the sequence would need to have a convergent subsequence which is not possible because every pair of terms is farther apart than $\varepsilon$. Thus for every $\varepsilon > 0$, there is an $\varepsilon$ net. Thus the metric space is totally bounded. Let $N_\varepsilon$ denote an $\varepsilon$ net. Let $D = \bigcup_{k=1}^{n} N_{1/2^k}$. Then this is a countable dense set. It is countable because it is the countable union of finite sets and it is dense because given a point, there is a point of $D$ within $1/2^k$ of it.

Also recall that a complete metric space is one for which every Cauchy sequence converges to a point in the metric space.

The following is the main theorem which relates these concepts.

Theorem 11.1.38  For $(X, d)$ a metric space, the following are equivalent.

1. $(X, d)$ is compact.
2. $(X, d)$ is sequentially compact.
3. $(X, d)$ is complete and totally bounded.

Proof: 1. $\Rightarrow$ 2. Let $\{x_n\}$ be a sequence. Suppose it fails to have a convergent subsequence. Then it follows right away that no value of the sequence is repeated infinitely often. If $\bigcup_{n=1}^{\infty} \{x_n\}$ has a limit point in $X$, then it follows from Proposition [Proposition 11.1.32] there would be a convergent subsequence converging to this limit point. Therefore, assume $\bigcup_{k=1}^{\infty} \{x_n\}$ has no limit point. This is equivalent to saying that $\bigcup_{k=m}^{\infty} \{x_k\}$ has no limit point for each $m$. Thus these are closed sets by Theorem [Theorem 11.1.34] because they contain all of their limit points due to the fact that they have none. Hence the open sets

$$\left(\bigcup_{k=m}^{\infty} \{x_n\}\right)^C$$

yield an open cover. This is an increasing sequence of open sets and none of them contain all the values of the sequence because no value is repeated for infinitely many indices. Thus this is an open cover which has no finite subcover contrary to 1.

2. $\Rightarrow$ 1. Suppose $(X, d)$ is sequentially compact. Then by Lemma [Lemma 11.1.37], $(X, d)$ is separable. Then it follows from Theorem [Theorem 11.1.32] that $(X, d)$ has the Lindelöf property. Suppose $(X, d)$ is not compact, then there is an open cover which admits no finite subcover $C$. Then by the Lindelöf property, there is a countable subcover of $C$ which also admits no subcover. Denote this by $\{U_i\}_{i=1}^{\infty}$. Then consider the non-empty closed sets

$$X \setminus \bigcup_{i=1}^{n} U_i \equiv X \cap \left(\bigcup_{i=1}^{n} U_i\right)^C \equiv F_n$$

Let $x_n \in F_n$. Then by assumption, there is a subsequence $\{x_{n_k}\}$ such that $\lim_{k \to \infty} x_{n_k} = x \in X$. Now for each $n$, $x \notin \bigcup_{i=1}^{n} U_i$ because if it were in $U_j$ for some $j \leq n$, then for all $k$ large enough, $x_{n_k} \in U_j$ which does not happen because in fact if $k > j$, $x_{n_k} \notin U_j$. Therefore, $\{U_i\}_{i=1}^{\infty}$ is not an open cover after all. It fails to cover $x$.

3. $\Rightarrow$ 2. Let $\{x_n\}$ be a sequence. Let $\{y_k\}_{k=1}^{\infty}$ be a $1/2^n$ net for $(X, d)$. Then $B\left(y_k, \frac{1}{2^n}\right)$ contains $x_n$ for infinitely many values of $n$. Out of the finitely many $\{y_k^{n+1}\}$ contained in $B\left(y_k, \frac{1}{2^n}\right)$, one
of these contains \( x_n \) for infinitely many values of \( n \). Denote this one by \( y_{k_n}^{n+1} \). Follow this process for \( n = 1, 2, \cdots \). Thus \( \{ y_{k_n}^n \}_{n=1}^{\infty} \) is a Cauchy sequence because for \( n > m \)

\[
d \left( y_{k_n}^n, y_{k_m}^m \right) \leq \sum_{l=m}^{n-1} d \left( y_{k_{l+1}}^{l+1}, y_{k_l}^l \right) \leq \sum_{l=m}^{\infty} \frac{1}{2^l} = \frac{1}{2^{m-1}}
\]

Thus, since \((X, d)\) is complete, there exists \( x \in X \) such that

\[
\lim_{n \to \infty} y_{k_n}^n = x
\]

Now also from the construction, there is a subsequence \( x_{k_n} \) such that \( d \left( x_{k_n}, y_{k_n}^n \right) < \frac{1}{2^n} \). It follows

\[
d \left( x, x_{k_n} \right) \leq d \left( x, y_{k_n}^n \right) + d \left( y_{k_n}^n, x_{k_n} \right) < d \left( x, y_{k_n}^n \right) + \frac{1}{2^n}
\]

The right side converges to 0 as \( n \to \infty \) and so \( \lim_{n \to \infty} d \left( x, x_{k_n} \right) = 0 \) also. Thus \( \{ x_n \} \) has a convergent subsequence after all.

2. \( \implies \) 3. If \((X, d)\) is sequentially compact, then by Lemma \( \implies \) it is totally bounded. If \( \{ x_n \} \) is a Cauchy sequence, then there is a subsequence which converges to \( x \in X \) by assumption. However, from Theorem \( \implies \) this requires the original Cauchy sequence to converge. \( \blacksquare \)

11.1.7 Lipschitz Continuity And Contraction Maps

The following is of more interest in the case of normed vector spaces, but there is no harm in stating it in this more general setting. You should verify that the functions described in the following definition are all continuous.

**Definition 11.1.39** Let \( f : X \to Y \) where \((X, d)\) and \((Y, \rho)\) are metric spaces. Then \( f \) is said to be **Lipschitz continuous** if for every \( x, \bar{x} \in X \), \( \rho (f(x), f(\bar{x})) \leq rd(x, \bar{x}) \). The function is called a **contraction map** if \( r < 1 \).

The big theorem about contraction maps is the following.

**Theorem 11.1.40** Let \( f : (X, d) \to (X, d) \) be a contraction map and let \((X, d)\) be a complete metric space. Thus Cauchy sequences converge and also \( d( f( x) , f( \bar{x}) ) \leq rd( x, \bar{x}) \) where \( r < 1 \). Then \( f \) has a unique fixed point. This is a point \( x \in X \) such that \( f(x) = x \). Also, if \( x_0 \) is any point of \( X \), then

\[
d \left( x, x_0 \right) \leq \frac{d \left( x_0, f(x_0) \right)}{1-r}
\]

Also, for each \( n \),

\[
d \left( f^n(x_0), x_0 \right) \leq \frac{d \left( x_0, f(x_0) \right)}{1-r},
\]

and \( x = \lim_{n \to \infty} f^n(x_0) \).

**Proof:** Pick \( x_0 \in X \) and consider the sequence of iterates of the map,

\[
x_0, f(x_0), f^2(x_0), \cdots.
\]

We argue that this is a Cauchy sequence. For \( m < n \), it follows from the triangle inequality,

\[
d \left( f^m(x_0), f^n(x_0) \right) \leq \sum_{k=m}^{n-1} d \left( f^{k+1}(x_0), f^k(x_0) \right) \leq \sum_{k=m}^{\infty} r^k d \left( f(x_0), x_0 \right)
\]

The reason for this last is as follows,

\[
d \left( f^2(x_0), f(x_0) \right) \leq rd \left( f(x_0), x_0 \right)
\]
11.1. METRIC SPACES

\[ d(f^3(x_0), f^2(x_0)) \leq rd(f^2(x_0), f(x_0)) \leq r^2d(f(x_0), x_0) \]

and so forth. Therefore,

\[ d(f^n(x_0), f^m(x_0)) \leq d(f(x_0), x_0) \frac{r^m}{1-r} \]

which shows that this is indeed a Cauchy sequence. Therefore, there exists \( x \) such that

\[ \lim_{n \to \infty} f^n(x_0) = x \]

By continuity,

\[ f(x) = f \left( \lim_{n \to \infty} f^n(x_0) \right) = \lim_{n \to \infty} f^{n+1}(x_0) = x. \]

Also note that this estimate yields

\[ d(x_0, f^n(x_0)) \leq \frac{d(x_0, f(x_0))}{1-r} \]

Now \( d(x_0, x) \leq d(x_0, f^n(x_0)) + d(f^n(x_0), x) \) and so

\[ d(x_0, x) - d(f^n(x_0), x) \leq \frac{d(x_0, f(x_0))}{1-r} \]

Letting \( n \to \infty \), it follows that

\[ d(x_0, x) \leq \frac{d(x_0, f(x_0))}{1-r} \]

It only remains to verify that there is only one fixed point. Suppose then that \( x, x' \) are two. Then

\[ d(x, x') = d(f(x), f(x')) \leq rd(x', x) \]

and so \( d(x, x') = 0 \) because \( r < 1 \). □

The above is the usual formulation of this important theorem, but we actually proved a better result.

**Corollary 11.1.41** Let \( B \) be a closed subset of the complete metric space \( (X, d) \) and let \( f : B \to X \) be a contraction map

\[ d(f(x), f(\hat{x})) \leq rd(x, \hat{x}), \ r < 1. \]

Also suppose there exists \( x_0 \in B \) such that the sequence of iterates \( \{f^n(x_0)\}_{n=1}^\infty \) remains in \( B \). Then \( f \) has a unique fixed point in \( B \) which is the limit of the sequence of iterates. This is a point \( x \in B \) such that \( f(x) = x \). In the case that \( B = \overline{B}(x_0, \delta) \), the sequence of iterates satisfies the inequality

\[ d(f^n(x_0), x_0) \leq \frac{d(x_0, f(x_0))}{1-r} \]

and so it will remain in \( B \) if

\[ \frac{d(x_0, f(x_0))}{1-r} < \delta. \]

**Proof:** By assumption, the sequence of iterates stays in \( B \). Then, as in the proof of the preceding theorem, for \( m < n \), it follows from the triangle inequality,

\[
\begin{align*}
d(f^n(x_0), f^m(x_0)) & \leq \sum_{k=m}^{n-1} d(f^{k+1}(x_0), f^k(x_0)) \\
& \leq \sum_{k=m}^\infty r^k d(f(x_0), x_0) = r^m d(f(x_0), x_0)
\end{align*}
\]
Hence the sequence of iterates is Cauchy and must converge to a point $x$ in $X$. However, $B$ is closed and so it must be the case that $x \in B$. Then as before,

$$x = \lim_{n \to \infty} f^n(x_0) = \lim_{n \to \infty} f^{n+1}(x_0) = f \left( \lim_{n \to \infty} f^n(x_0) \right) = f(x)$$

As to the sequence of iterates remaining in $B$ where $B$ is a ball as described, the inequality above in the case where $m = 0$ yields

$$d(x_0, f^n(x_0)) \leq \frac{1}{1-r} d(f(x_0), x_0)$$

and so, if the right side is less than $\delta$, then the iterates remain in $B$. As to the fixed point being unique, it is as before. If $x, x'$ are both fixed points in $B$, then $d(x, x') = d(f(x), f(x')) \leq rd(x, x')$ and so $x = x'$.

The contraction mapping theorem has an extremely useful generalization. In order to get a unique fixed point, it suffices to have some power of $f$ a contraction map.

**Theorem 11.1.42** Let $f : (X, d) \to (X, d)$ have the property that for some $n \in \mathbb{N}$, $f^n$ is a contraction map and let $(X, d)$ be a complete metric space. Then there is a unique fixed point for $f$. As in the earlier theorem the sequence of iterates $\{f^n(x_0)\}_{n=1}^{\infty}$ also converges to the fixed point.

**Proof:** From Theorem 11.1.40 there is a unique fixed point for $f^n$. Thus

$$f^n(x) = x$$

Then

$$f^n(f(x)) = f^{n+1}(x) = f(x)$$

By uniqueness, $f(x) = x$.

Now consider the sequence of iterates. Suppose it fails to converge to $x$. Then there is $\varepsilon > 0$ and a subsequence $n_k$ such that

$$d(f^{n_k}(x_0), x) \geq \varepsilon$$

Now $n_k = p_k n + r_k$ where $r_k$ is one of the numbers $\{0, 1, 2, \ldots, n-1\}$. It follows that there exists one of these numbers which is repeated infinitely often. Call it $r$ and let the further subsequence continue to be denoted as $n_k$. Thus

$$d(f^{p_k n + r}(x_0), x) \geq \varepsilon$$

In other words,

$$d(f^{p_k n}(f^r(x_0)), x) \geq \varepsilon$$

However, from Theorem 11.1.40, as $k \to \infty$, $f^{p_k n}(f^r(x_0)) \to x$ which contradicts the above inequality. Hence the sequence of iterates converges to $x$, as it did for $f$ a contraction map.

Now with the above material on analysis, it is time to begin using the ideas from linear algebra in this special case where the field of scalars is $\mathbb{R}$ or $\mathbb{C}$.

### 11.1.8 Convergence Of Functions

Next is to consider the meaning of convergence of sequences of functions. There are two main ways of convergence of interest here, pointwise and uniform convergence.

**Definition 11.1.43** Let $f_n : X \to Y$ where $(X, d), (Y, \rho)$ are two metric spaces. Then $\{f_n\}$ is said to converge pointwise to a function $f : X \to Y$ if for every $x \in X$,

$$\lim_{n \to \infty} f_n(x) = f(x)$$

$\{f_n\}$ is said to converge uniformly if for all $\varepsilon > 0$, there exists $N$ such that if $n \geq N$, then

$$\sup_{x \in X} \rho(f_n(x), f(x)) < \varepsilon$$
11.2. CONNECTED SETS

Here is a well known example illustrating the difference between pointwise and uniform convergence.

**Example 11.1.44** Let \( f_n (x) = x^n \) on the metric space \([0, 1]\). Then this function converges pointwise to

\[
f(x) = \begin{cases} 
0 & \text{on } [0, 1) \\
1 & \text{at } 1 
\end{cases}
\]

but it does not converge uniformly on this interval to \( f \).

Note how the target function \( f \) in the above example is not continuous even though each function in the sequence is. The nice thing about uniform convergence is that it takes continuity of the functions in the sequence and imparts it to the target function. It does this for both continuity at a single point and uniform continuity. Thus uniform convergence is a very superior thing.

**Theorem 11.1.45** Let \( f_n : X \to Y \) where \((X, d), (Y, \rho)\) are two metric spaces and suppose each \( f_n \) is continuous at \( x \in X \) and also that \( f_n \) converges uniformly to \( f \) on \( X \). Then \( f \) is also continuous at \( x \). In addition to this, if each \( f_n \) is uniformly continuous on \( X \), then the same is true for \( f \).

**Proof:** Let \( \varepsilon > 0 \) be given. Then

\[
\rho(f(x), f(\hat{x})) \leq \rho(f(x), f_n(x)) + \rho(f_n(x), f_N(\hat{x})) + \rho(f_N(\hat{x}), f(\hat{x}))
\]

By uniform convergence, there exists \( N \) such that both \( \rho(f(x), f_n(x)) \) and \( \rho(f_n(\hat{x}), f(\hat{x})) \) are less than \( \varepsilon/3 \) provided \( n \geq N \). Thus picking such an \( n \),

\[
\rho(f(x), f(\hat{x})) \leq \frac{2\varepsilon}{3} + \rho(f_n(x), f_n(\hat{x}))
\]

Now from the continuity of \( f_n \), there exists \( \delta > 0 \) such that if \( d(x, \hat{x}) < \delta \), then \( \rho(f_n(x), f_n(\hat{x})) < \varepsilon/3 \). Hence, if \( d(x, \hat{x}) < \delta \), then

\[
\rho(f(x), f(\hat{x})) \leq \frac{2\varepsilon}{3} + \rho(f_n(x), f_n(\hat{x})) < \frac{2\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon
\]

Hence, \( f \) is continuous at \( x \).

Next consider uniform continuity. It follows from the uniform convergence that if \( x, \hat{x} \) are any two points of \( X \), then if \( n \geq N \), then, picking such an \( n \),

\[
\rho(f(x), f(\hat{x})) \leq \frac{2\varepsilon}{3} + \rho(f_n(x), f_n(\hat{x}))
\]

By uniform continuity of \( f_n \) there exists \( \delta \) such that if \( d(x, \hat{x}) < \delta \), then the term on the right in the above is less than \( \varepsilon/3 \). Hence if \( d(x, \hat{x}) < \delta \), then \( \rho(f(x), f(\hat{x})) < \varepsilon \) and so \( f \) is uniformly continuous as claimed. \( \blacksquare \)

### 11.2 Connected Sets

This has absolutely nothing to do with linear algebra but is here to provide convenient results to be used later when linear algebra will occur as part of some topics in analysis.

Stated informally, connected sets are those which are in one piece. In order to define what is meant by this, I will first consider what it means for a set to **not** be in one piece. This is called **separated**. Connected sets are defined in terms of **not** being separated. This is why theorems about connected sets sometimes seem a little tricky.

**Definition 11.2.1** A set, \( S \) in a metric space, is separated if there exist sets \( A, B \) such that

\[
S = A \cup B, \ A, B \neq \emptyset, \text{ and } \overline{A} \cap B = B \cap \overline{A} = \emptyset.
\]

In this case, the sets \( A \) and \( B \) are said to separate \( S \). A set is connected if it is not separated. Remember \( \overline{A} \) denotes the closure of the set \( A \).
Note that the concept of connected sets is defined in terms of what it is not. This makes it somewhat difficult to understand. One of the most important theorems about connected sets is the following.

**Theorem 11.2.2** Suppose $\mathcal{U}$ is a set of connected sets and that there exists a point $p$ which is in all of these connected sets. Then $K \equiv \bigcup \mathcal{U}$ is connected.

**Proof:** Suppose

$$K = A \cup B$$

where $\bar{A} \cap B = \bar{B} \cap A = \emptyset$, $A \neq \emptyset$, $B \neq \emptyset$. Let $U \in \mathcal{U}$. Then

$$U = (U \cap A) \cup (U \cap B)$$

and this would separate $U$ if both sets in the union are nonempty since the limit points of $U \cap B$ are contained in the limit points of $B$. It follows that every set of $\mathcal{U}$ is contained in one of $A$ or $B$. Suppose then that some $U \subseteq A$. Then all $U \in \mathcal{U}$ must be contained in $A$ because if one is contained in $B$, this would violate the assumption that they all have a point $p$ in common. Thus $K$ is connected after all because this requires $B = \emptyset$. Alternatively, $p$ is in one of these sets. Say $p \in A$. Then by the above argument every $U$ must be in $A$ because if not, the above would be a separation of $U$. Thus $B = \emptyset$. ■

The intersection of connected sets is not necessarily connected as is shown by the following picture.

![Connected Component Diagram](image)

**Theorem 11.2.3** Let $f : X \to Y$ be continuous where $Y$ is a metric space and $X$ is connected. Then $f(X)$ is also connected.

**Proof:** To do this you show $f(X)$ is not separated. Suppose to the contrary that $f(X) = A \cup B$ where $A$ and $B$ separate $f(X)$. Then consider the sets $f^{-1}(A)$ and $f^{-1}(B)$. If $z \in f^{-1}(B)$, then $f(z) \in B$ and so $f(z)$ is not a limit point of $A$. Therefore, there exists an open set, $U$ containing $f(z)$ such that $U \cap A = \emptyset$. But then, the continuity of $f$ and Theorem 11.2.3 implies that $f^{-1}(U)$ is an open set containing $z$ such that $f^{-1}(U) \cap f^{-1}(A) = \emptyset$. Therefore, $f^{-1}(B)$ contains no limit points of $f^{-1}(A)$. Similar reasoning implies $f^{-1}(A)$ contains no limit points of $f^{-1}(B)$. It follows that $X$ is separated by $f^{-1}(A)$ and $f^{-1}(B)$, contradicting the assumption that $X$ was connected. ■

An arbitrary set can be written as a union of maximal connected sets called connected components. This is the concept of the next definition.

**Definition 11.2.4** Let $S$ be a set and let $p \in S$. Denote by $C_p$ the union of all connected subsets of $S$ which contain $p$. This is called the connected component determined by $p$.

**Theorem 11.2.5** Let $C_p$ be a connected component of a set $S$ in a metric space. Then $C_p$ is a connected set and if $C_p \cap C_q \neq \emptyset$, then $C_p = C_q$.

**Proof:** Let $\mathcal{C}$ denote the connected subsets of $S$ which contain $p$. By Theorem 11.2.3, $\bigcup \mathcal{C} = C_p$ is connected. If $x \in C_p \cap C_q$, then from Theorem 11.2.4, $C_p \supseteq C_p \cup C_q$ and so $C_p \supseteq C_q$. The inclusion goes the other way by the same reason. ■

This shows the connected components of a set are equivalence classes and partition the set.

A set, $I$ is an interval in $\mathbb{R}$ if and only if whenever $x, y \in I$ then $(x, y) \subseteq I$. The following theorem is about the connected sets in $\mathbb{R}$.
11.2. CONNECTED SETS

**Theorem 11.2.6** A set $C$ in $\mathbb{R}$ is connected if and only if $C$ is an interval.

**Proof:** Let $C$ be connected. If $C$ consists of a single point, $p$, there is nothing to prove. The interval is just $[p,p]$. Suppose $p < q$ and $p,q \in C$. You need to show $(p,q) \subseteq C$. If

$$x \in (p,q) \setminus C$$

let $C \cap (-\infty,x) \equiv A$, and $C \cap (x,\infty) \equiv B$. Then $C = A \cup B$ and the sets $A$ and $B$ separate $C$ contrary to the assumption that $C$ is connected.

Conversely, let $I$ be an interval. Suppose $I$ is separated by $A$ and $B$. Pick $x \in A$ and $y \in B$. Suppose without loss of generality that $x < y$. Now define the set,

$$S \equiv \{ t \in [x,y] : [x,t] \subseteq A \}$$

and let $l$ be the least upper bound of $S$. Then $l \in A$ so $l \notin B$ which implies $l \in A$. But if $l \notin B$, then for some $\delta > 0$,

$$(l, l + \delta) \cap B = \emptyset$$

contradicting the definition of $l$ as an upper bound for $S$. Therefore, $l \in B$ which implies $l \notin A$ after all, a contradiction. It follows $I$ must be connected. $\blacksquare$

This yields a generalization of the intermediate value theorem from one variable calculus.

**Corollary 11.2.7** Let $E$ be a connected set in a metric space and suppose $f : E \to \mathbb{R}$ and that $y \in (f(e_1), f(e_2))$ where $e_i \in E$. Then there exists $e \in E$ such that $f(e) = y$.

**Proof:** From Theorem 11.2.3, $f(E)$ is a connected subset of $\mathbb{R}$. By Theorem 11.2.6, $f(E)$ must be an interval. In particular, it must contain $y$. This proves the corollary. $\blacksquare$

The following theorem is a very useful description of the open sets in $\mathbb{R}$.

**Theorem 11.2.8** Let $U$ be an open set in $\mathbb{R}$. Then there exist countably many disjoint open sets \{$(a_i, b_i)$\}_{i=1}^{\infty} such that $U = \bigcup_{i=1}^{\infty} (a_i, b_i)$.

**Proof:** Let $p \in U$ and let $z \in C_p$, the connected component determined by $p$. Since $U$ is open, there exists, $\delta > 0$ such that $(z - \delta, z + \delta) \subseteq U$. It follows from Theorem 11.2.2 that

$$(z - \delta, z + \delta) \subseteq C_p.$$ 

This shows $C_p$ is open. By Theorem 11.2.7, this shows $C_p$ is an open interval, $(a,b)$ where $a,b \in [ -\infty, \infty ]$. There are therefore at most countably many of these connected components because each must contain a rational number and the rational numbers are countable. Denote by \{$(a_i, b_i)$\}_{i=1}^{\infty} the set of these connected components. $\blacksquare$

**Definition 11.2.9** A set $E$ in a metric space is arcwise connected if for any two points, $p, q \in E$, there exists a closed interval, $[a, b]$ and a continuous function, $\gamma : [a, b] \to E$ such that $\gamma (a) = p$ and $\gamma (b) = q$.

An example of an arcwise connected metric space would be any subset of $\mathbb{R}^n$ which is the continuous image of an interval. Arcwise connected is not the same as connected. A well known example is the following.

\[
\left\{ \left( x, \sin \frac{1}{x} \right) : x \in (0,1) \right\} \cup \left\{ (0, y) : y \in [-1,1] \right\}
\]

(11.2)

You can verify that this set of points in the normed vector space $\mathbb{R}^2$ is not arcwise connected but is connected.
11.3 Subspaces, Spans and Bases

As shown earlier, \( \mathbb{F}^n \) is an example of a vector space with field of scalars \( \mathbb{F} \). Here is a short review of the major exchange theorem. Here and elsewhere, when it is desired to emphasize that certain things are vectors, bold face will be used. However, sometimes the context makes this sufficiently clear and bold face is not used.

**Theorem 11.3.1** If
\[
\text{span}\left(u_1, \ldots, u_r\right) \subseteq \text{span}\left(v_1, \ldots, v_s\right) \equiv V
\]
and \( \{u_1, \ldots, u_r\} \) are linearly independent, then \( r \leq s \).

**Proof:** Suppose \( r > s \). Let \( E_p \) denote a finite list of vectors of \( \{v_1, \ldots, v_s\} \) and let \( |E_p| \) denote the number of vectors in the list. Let \( F_p \) denote the first \( p \) vectors in \( \{u_1, \ldots, u_r\} \). In case \( p = 0 \), \( F_p \) will denote the empty set. For \( 0 \leq p \leq s \), let \( E_p \) have the property
\[
\text{span}\left(F_p, E_p\right) = V
\]
and \( |E_p| \) is as small as possible for this to happen. I claim \( |E_p| \leq s - p \) if \( E_p \) is nonempty.

Here is why. For \( p = 0 \), it is obvious. Suppose true for some \( p < s \). Then
\[
u_{p+1} \in \text{span}\left(F_p, E_p\right)
\]
and so there are constants, \( c_1, \ldots, c_p \) and \( d_1, \ldots, d_m \) where \( m \leq s - p \) such that
\[
u_{p+1} = \sum_{i=1}^{p} c_i u_i + \sum_{j=1}^{m} d_j z_j
\]
for
\[\{z_1, \ldots, z_m\} \subseteq \{v_1, \ldots, v_s\}.\]

Then not all the \( d_j \) can equal zero because this would violate the linear independence of the \( \{u_1, \ldots, u_r\} \). Therefore, you can solve for one of the \( z_k \) as a linear combination of \( \{u_1, \ldots, u_{p+1}\} \) and the other \( z_j \). Thus you can change \( F_p \) to \( F_{p+1} \) and include one fewer vector in \( E_p \). Thus
\[|E_{p+1}| \leq m - 1 \leq s - p - 1.\]
This proves the claim.

Therefore, \( E_s \) is empty and \( \text{span}\left(u_1, \ldots, u_s\right) = V \). However, this gives a contradiction because it would require
\[
u_{s+1} \in \text{span}\left(u_1, \ldots, u_s\right)
\]
which violates the linear independence of these vectors. \( \blacksquare \)

Also recall the following.

**Definition 11.3.2** A finite set of vectors, \( \{x_1, \ldots, x_r\} \) is a basis for a vector space \( V \) if
\[
\text{span}\left(x_1, \ldots, x_r\right) = V
\]
and \( \{x_1, \ldots, x_r\} \) is linearly independent. Thus if \( v \in V \) there exist unique scalars, \( v_1, \ldots, v_r \) such that \( v = \sum_{i=1}^{r} v_i x_i \). These scalars are called the components of \( v \) with respect to the basis \( \{x_1, \ldots, x_r\} \).

**Corollary 11.3.3** Let \( \{x_1, \ldots, x_r\} \) and \( \{y_1, \ldots, y_s\} \) be two bases\(^3\) of \( \mathbb{F}^n \). Then \( r = s = n \).

**Lemma 11.3.4** Let \( \{v_1, \ldots, v_r\} \) be a set of vectors. Then \( V \equiv \text{span}\left(v_1, \ldots, v_r\right) \) is a subspace.

**Definition 11.3.5** Let \( V \) be a vector space. Then \( \dim(V) \) read as the dimension of \( V \) is the number of vectors in a basis.

\(^3\)This is the plural form of basis. We could say basiss but it would involve an inordinate amount of hissing as in “The sixth shiek’s sixth sheep is sick”. This is the reason that bases is used instead of basiss.
Of course you should wonder right now whether an arbitrary subspace of a finite dimensional vector space even has a basis. In fact it does and this is in the next theorem. First, here is an interesting lemma which was also presented earlier.

**Lemma 11.3.6** Suppose \( v \notin \text{span}(u_1, \cdots, u_k) \) and \( \{u_1, \cdots, u_k, v\} \) is also linearly independent.

**Theorem 11.3.7** Let \( V \) be a nonzero subspace of \( Y \) a finite dimensional vector space having dimension \( n \). Then \( V \) has a basis.

In words the following corollary states that any linearly independent set of vectors can be enlarged to form a basis.

**Corollary 11.3.8** Let \( V \) be a subspace of \( Y \), a finite dimensional vector space of dimension \( n \) and let \( \{v_1, \cdots, v_r\} \) be a linearly independent set of vectors in \( V \). Then either it is a basis for \( V \) or there exist vectors, \( v_{r+1}, \cdots, v_s \) such that \( \{v_1, \cdots, v_r, v_{r+1}, \cdots, v_s\} \) is a basis for \( V \).

**Theorem 11.3.9** Let \( V \) be a subspace of \( Y \), a finite dimensional vector space of dimension \( n \) and suppose \( \text{span}(u_1, \cdots, u_p) = V \) where the \( u_i \) are nonzero vectors. Then there exist vectors, \( \{v_1, \cdots, v_r\} \) such that \( \{v_1, \cdots, v_r\} \subseteq \{u_1, \cdots, u_p\} \) and \( \{v_1, \cdots, v_r\} \) is a basis for \( V \).

### 11.4 Inner Product and Normed Linear Spaces

#### 11.4.1 The Inner Product In \( \mathbb{F}^n \)

To do calculus, you must understand what you mean by distance. For functions of one variable, the distance was provided by the absolute value of the difference of two numbers. This must be generalized to \( \mathbb{F}^n \) and to more general situations. This is the most familiar setting for elementary courses. We call it the dot product in calculus and physics but it is a case of something which also works in \( \mathbb{C}^n \).

**Definition 11.4.1** Let \( x, y \in \mathbb{F}^n \). Thus \( x = (x_1, \cdots, x_n) \) where each \( x_k \in \mathbb{F} \) and a similar formula holding for \( y \). Then the inner product of these two vectors is defined to be

\[
(x \cdot y) = (x, y) = \sum_j x_j \overline{y_j} = x_1 \overline{y_1} + \cdots + x_n \overline{y_n}.
\]

This is also often denoted by \((x, y)\) or as \( \langle x, y \rangle \) and is called an inner product. I will use either notation.

Notice how you put the conjugate on the entries of the vector, \( y \). It makes no difference if the vectors happen to be real vectors but with complex vectors you must do it this way\(^2\). The reason for this is that when you take the inner product of a vector with itself, you want to get the square of the length of the vector, a positive number. Placing the conjugate on the components of \( y \) in the above definition assures this will take place. Thus

\[
(x, x) = \sum_j x_j \overline{x_j} = \sum_j |x_j|^2 \geq 0.
\]

If you didn’t place a conjugate as in the above definition, things wouldn’t work out correctly. For example,

\[
(1 + i)^2 + 2^2 = 4 + 2i
\]

\(^2\)Sometimes people put the conjugate on the components of the first entry. It doesn’t matter a lot, but it is good to be consistent. I have chosen to place the conjugate on the components of the second entry.
and this is not a positive number.

The following properties of the inner product follow immediately from the definition and you should verify each of them.

Properties of the inner product:

1. \((u, v) = (v, u)\)

2. If \(a, b\) are numbers and \(u, v, z\) are vectors then \(((au + bv), z) = a(u, z) + b(v, z)\).

3. \((u, u) \geq 0\) and it equals 0 if and only if \(u = 0\).

Note this implies \((x, \alpha y) = \bar{\alpha} (x, y)\) because

\[(x, \alpha y) = (\alpha y, x) = \bar{\alpha} (y, x) = \bar{\alpha} (x, y)\]

The norm is defined as follows.

**Definition 11.4.2** For \(x \in \mathbb{F}^n\),

\[|x| = \left( \sum_{k=1}^{n} |x_k|^2 \right)^{1/2} = (x, x)^{1/2}\]

**11.4.2 General Inner Product Spaces**

Any time you have a vector space which possesses an inner product, something satisfying the properties 1-3 above, it is called an inner product space.

Here is a fundamental inequality called the **Cauchy Schwarz inequality** which holds in any inner product space. First here is a simple lemma.

**Lemma 11.4.3** If \(z \in \mathbb{F}\) there exists \(\zeta \in \mathbb{F}\) such that \(|z| = |\zeta|, |\zeta| = 1\).

**Proof:** Let \(\theta = 1\) if \(z = 0\) and otherwise, let \(\theta = \frac{\bar{\zeta}}{|z|}\). Recall that for \(z = x + iy, \bar{\zeta} = x - iy\) and \(\bar{\zeta} z = |z|^2\). In case \(z\) is real, there is no change in the above. \(\Box\)

**Theorem 11.4.4** (Cauchy Schwarz) Let \(H\) be an inner product space. The following inequality holds for \(x\) and \(y \in H\):

\[|(x, y)| \leq (x, x)^{1/2} (y, y)^{1/2}\] \hspace{1cm} (11.3)

Equality holds in this inequality if and only if one vector is a multiple of the other.

**Proof:** Let \(\theta \in \mathbb{F}\) such that \(|\theta| = 1\) and

\[\theta (x, y) = |(x, y)|\]

Consider \(p(t) \equiv (x + \bar{\theta} t y, x + t \bar{y} y)\) where \(t \in \mathbb{R}\). Then from the above list of properties of the inner product,

\[
0 \leq p(t) = (x, x) + t \theta (x, y) + t \bar{\theta} (y, x) + t^2 (y, y) \\
= (x, x) + t \theta (x, y) + t \bar{\theta} (x, y) + t^2 (y, y) \\
= (x, x) + 2t \Re (\theta (x, y)) + t^2 (y, y) \\
= (x, x) + 2t |(x, y)| + t^2 (y, y) \hspace{1cm} (11.4)
\]

and this must hold for all \(t \in \mathbb{R}\). Therefore, if \((y, y) = 0\) it must be the case that \(|(x, y)| = 0\) also since otherwise the above inequality would be violated. Therefore, in this case,

\[|(x, y)| \leq (x, x)^{1/2} (y, y)^{1/2}\]
On the other hand, if \((y, y) \neq 0\), then \(p(t) \geq 0\) for all \(t\) means the graph of \(y = p(t)\) is a parabola which opens up and it either has exactly one real zero in the case its vertex touches the \(t\) axis or it has no real zeros. From the quadratic formula this happens exactly when
\[
4 |(x, y)|^2 - 4 (x, x) (y, y) \leq 0
\]
which is equivalent to \(\|x\| = |x|^{1/2}\).

It is clear from a computation that if one vector is a scalar multiple of the other that equality holds in 11.3. Conversely, suppose equality does hold. Then this is equivalent to saying \(4 |(x, y)|^2 - 4 (x, x) (y, y) = 0\) and so from the quadratic formula, there exists one real zero to \(p(t) = 0\). Call it \(t_0\). Then
\[
p(t_0) = (x + \bar{t}_0 y, x + t_0 \bar{y}) = |x + \bar{t}_0 y|^2 = 0
\]
and so \(x = -\bar{t}_0 y\). This proves the theorem. □

Note that in establishing the inequality, I only used part of the above properties of the inner product. It was not necessary to use the one which says that if \((x, x) = 0\) then \(x = 0\).

Now the length of a vector can be defined.

**Definition 11.4.5** Let \(z \in H\). Then \(|z| \equiv (z, z)^{1/2}\).

**Theorem 11.4.6** For length defined in Definition 11.4.5, the following hold.

\[
|z| \geq 0 \quad \text{and} \quad |z| = 0 \quad \text{if and only if} \quad z = 0
\]

If \(\alpha\) is a scalar, \(|\alpha z| = |\alpha| |z|\)

\[
|z + w| \leq |z| + |w|.
\]

**Proof:** The first two claims are left as exercises. To establish the third,
\[
|z + w|^2 \equiv (z + w, z + w) = (z, z) + (w, w) + (w, z) + (z, w)
\]
\[
= |z|^2 + |w|^2 + 2 \text{Re}(w, z)
\]
\[
\leq |z|^2 + |w|^2 + 2 |w| |z| = (|z| + |w|)^2.
\]

One defines the distance between two vectors \(x, y\) in an inner product space as \(|x - y|\).

### 11.4.3 Normed Vector Spaces

The best sort of a norm is one which comes from an inner product, because these norms preserve familiar geometrical ideas. However, any vector space, \(V\) which has a function, \(||\cdot||\) which maps \(V\) to \([0, \infty)\) is called a normed vector space if \(||\cdot||\) satisfies \(\|z\| \leq \|x\| + \|w\|\). That is
\[
||z|| \geq 0 \quad \text{and} \quad ||z|| = 0 \quad \text{if and only if} \quad z = 0
\]

If \(\alpha\) is a scalar, \(||\alpha z|| = |\alpha||z||\)

\[
||z + w|| \leq ||z|| + ||w||.
\]

The last inequality above is called the triangle inequality. Another version of this is
\[
||z|| - ||w|| \leq ||z - w||
\]

To see that \(\|z\| \leq \|z - w + w\| \leq \|z - w\| + \|w\|\)
which implies
\[ |z| - |w| \leq |z - w| \]
and now switching \( z \) and \( w \), yields
\[ |w| - |z| \leq |z - w| \]
which implies \( ||w|| - ||z|| \leq |z - w| \)

The distance between \( x, y \) is given by
\[ \|x - y\| \]
This distance satisfies
\[ \|x - y\| = \|y - x\| \]
\[ \|x - y\| \geq 0 \text{ and is } 0 \text{ if and only if } x = y \]
\[ \|x - y\| \leq \|x - z\| + \|z - y\| \]
Thus this yields a metric space, but it has more because it also involves interaction with the algebra of the vector space.

11.4.4 The \( p \) Norms

Examples of norms are the \( p \) norms on \( \mathbb{C}^n \). These do not come from an inner product but they are norms just the same.

**Definition 11.4.7** Let \( x \in \mathbb{C}^n \). Then define for \( p \geq 1 \),
\[ ||x||_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p} \]

The following inequality is called Holder’s inequality.

**Proposition 11.4.8** For \( x, y \in \mathbb{C}^n \),
\[ \sum_{i=1}^{n} |x_i| |y_i| \leq \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p} \left( \sum_{i=1}^{n} |y_i|^{p'} \right)^{1/p'} \]

The proof will depend on the following lemma.

**Lemma 11.4.9** If \( a, b \geq 0 \) and \( p' \) is defined by \( \frac{1}{p} + \frac{1}{p'} = 1 \), then
\[ ab \leq \frac{a^p}{p} + \frac{b^{p'}}{p'} \]

**Proof of the Proposition:** If \( x \) or \( y \) equals the zero vector there is nothing to prove. Therefore, assume they are both nonzero. Let \( A = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p} \) and \( B = \left( \sum_{i=1}^{n} |y_i|^{p'} \right)^{1/p'} \). Then using Lemma 11.4.9,
\[ \sum_{i=1}^{n} \frac{|x_i| |y_i|}{A B} \leq \sum_{i=1}^{n} \left[ \frac{1}{p} \left( \frac{|x_i|}{A} \right)^p \frac{1}{p'} \left( \frac{|y_i|}{B} \right)^{p'} \right] \]
\[ = \frac{1}{p} \frac{1}{A^p} \sum_{i=1}^{n} |x_i|^p + \frac{1}{p'} \frac{1}{B^{p'}} \sum_{i=1}^{n} |y_i|^{p'} \]
\[ = \frac{1}{p} + \frac{1}{p'} = 1 \]
and so
\[ \sum_{i=1}^{n} |x_i| |y_i| \leq AB = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p} \left( \sum_{i=1}^{n} |y_i|^{p'} \right)^{1/p'} . \]
Theorem 11.4.10 The $p$ norms do indeed satisfy the axioms of a norm.

Proof: It is obvious that $||\cdot||_p$ does indeed satisfy most of the norm axioms. The only one that is not clear is the triangle inequality. To save notation write $||\cdot||$ in place of $||\cdot||_p$ in what follows. Note also that $p = 1 + \frac{1}{p'}$.

Then using the Holder inequality,

$$||x + y||_p^p = \sum_{i=1}^{n} |x_i + y_i|^p$$

$$\leq \sum_{i=1}^{n} |x_i + y_i|^{p-1} |x_i| + \sum_{i=1}^{n} |x_i + y_i|^{p-1} |y_i|$$

$$= \sum_{i=1}^{n} |x_i + y_i|^p |x_i| + \sum_{i=1}^{n} |x_i + y_i|^p |y_i|$$

$$\leq \left( \sum_{i=1}^{n} |x_i + y_i|^p \right)^{1/p'} \left[ \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p} + \left( \sum_{i=1}^{n} |y_i|^p \right)^{1/p'} \right]$$

$$= ||x + y||^{p/p'} \left( ||x||_p + ||y||_p \right)$$

so dividing by $||x + y||^{p/p'}$, it follows

$$||x + y||^p ||x + y||^{-p/p'} = ||x + y|| \leq ||x||_p + ||y||_p$$

$$\left( p - \frac{p}{p'} = p \left( 1 - \frac{1}{p'} \right) = p \frac{1}{p} = 1 \right).$$

It only remains to prove Lemma 11.4.9.

Proof of the lemma: Let $p' = q$ to save on notation and consider the following picture:

Note equality occurs when $a^p = b^q$.

Alternate proof of the lemma: First note that if either $a$ or $b$ are zero, then there is nothing to show so we can assume $b, a > 0$. Let $b > 0$ and let

$$f(a) = \frac{a^p}{p} + \frac{b^q}{q} - ab$$

Then the second derivative of $f$ is positive on $(0, \infty)$ so its graph is convex. Also $f(0) > 0$ and $\lim_{a \to \infty} f(a) = \infty$. Then a short computation shows that there is only one critical point, where $f$ is minimized and this happens when $a$ is such that $a^p = b^q$. At this point,

$$f(a) = b^q - b^{q/p} b = b^q - b^{q-1} b = 0$$

Therefore, $f(a) \geq 0$ for all $a$ and this proves the lemma.
Another example of a very useful norm on $F^n$ is the norm $\| \cdot \|_\infty$ defined by

$$\| x \|_\infty \equiv \max \{ |x_k| : k = 1, 2, \cdots, n \}$$

You should verify that this satisfies all the axioms of a norm. Here is the triangle inequality.

$$\| x + y \|_\infty = \max_k \{ |x_k + y_k| \} \leq \max_k \{ |x_k| + |y_k| \} \leq \max_k \{ |x_k| \} + \max_k \{ |y_k| \} = \| x \|_\infty + \| y \|_\infty$$

It turns out that in terms of analysis (limits of sequences, completeness and so forth), it makes absolutely no difference which norm you use. There are however, significant geometric differences. This will be explained later. First is the notion of an orthonormal basis.

### 11.4.5 Orthonormal Bases

Not all bases for an inner product space $H$ are created equal. The best bases are orthonormal.

**Definition 11.4.11** Suppose $\{v_1, \cdots, v_k\}$ is a set of vectors in an inner product space $H$. It is an orthonormal set if

$$(v_i, v_j) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Every orthonormal set of vectors is automatically linearly independent.

**Proposition 11.4.12** Suppose $\{v_1, \cdots, v_k\}$ is an orthonormal set of vectors. Then it is linearly independent.

**Proof:** Suppose $\sum_{i=1}^k c_i v_i = 0$. Then taking inner products with $v_j$,

$$0 = (0, v_j) = \sum_i c_i (v_i, v_j) = \sum_i c_i \delta_{ij} = c_j.$$

Since $j$ is arbitrary, this shows the set is linearly independent as claimed. □

It turns out that if $X$ is any subspace of $H$, then there exists an orthonormal basis for $X$.

**Lemma 11.4.13** Let $X$ be a subspace of dimension $n$ whose basis is $\{x_1, \cdots, x_n\}$. Then there exists an orthonormal basis for $X$, $\{u_1, \cdots, u_n\}$ which has the property that for each $k \leq n$, $\text{span}(x_1, \cdots, x_k) = \text{span}(u_1, \cdots, u_k)$.

**Proof:** Let $\{x_1, \cdots, x_n\}$ be a basis for $X$. Let $u_1 \equiv x_1 / |x_1|$. Thus for $k = 1$, $\text{span}(u_1) = \text{span}(x_1)$ and $\{u_1\}$ is an orthonormal set. Now suppose for some $k < n$, $u_1, \cdots, u_k$ have been chosen such that $(u_j, u_i) = \delta_{ij}$ and $\text{span}(x_1, \cdots, x_k) = \text{span}(u_1, \cdots, u_k)$. Then define

$$u_{k+1} \equiv \frac{x_{k+1} - \sum_{j=1}^k (x_{k+1}, u_j) u_j}{\| x_{k+1} - \sum_{j=1}^k (x_{k+1}, u_j) u_j \|}, \quad (11.12)$$

where the denominator is not equal to zero because the $x_j$ form a basis and so

$$x_{k+1} \notin \text{span}(x_1, \cdots, x_k) = \text{span}(u_1, \cdots, u_k)$$

Thus by induction,

$$u_{k+1} \in \text{span}(u_1, \cdots, u_k, x_{k+1}) = \text{span}(x_1, \cdots, x_k, x_{k+1})$$

Also, $x_{k+1} \in \text{span}(u_1, \cdots, u_k, u_{k+1})$ which is seen easily by solving $\Box$ for $x_{k+1}$ and it follows

$$\text{span}(x_1, \cdots, x_k, x_{k+1}) = \text{span}(u_1, \cdots, u_k, u_{k+1}).$$
If \( l \leq k \), then denoting by \( C \) the scalar \( \left| x_{k+1} - \sum_{j=1}^{k} (x_{k+1}, u_j) u_j \right|^{-1} \),

\[
(u_{k+1}, u_l) = C \left( (x_{k+1}, u_l) - \sum_{j=1}^{k} (x_{k+1}, u_j) (u_j, u_l) \right)
\]

\[
= C \left( (x_{k+1}, u_l) - \sum_{j=1}^{k} (x_{k+1}, u_j) \delta_{ij} \right)
\]

\[
= C ((x_{k+1}, u_l) - (x_{k+1}, u_l)) = 0.
\]

The vectors, \( \{u_j\}_{j=1}^{n-1} \), generated in this way are therefore an orthonormal basis because each vector has unit length. □

The process by which these vectors were generated is called the Gram Schmidt process.

The following corollary is obtained from the above process.

**Corollary 11.4.14** Let \( X \) be a finite dimensional inner product space of dimension \( n \) whose basis is \( \{u_1, \cdots, u_k, x_{k+1}, \cdots, x_n\} \). Then if \( \{u_1, \cdots, u_k\} \) is orthonormal, then the Gram Schmidt process applied to the given list of vectors in order leaves \( \{u_1, \cdots, u_k\} \) unchanged.

### 11.5 Equivalence Of Norms

As mentioned above, it makes absolutely no difference which norm you decide to use. This holds in general finite dimensional normed spaces and is shown here.

The fundamental lemma is the following which relates convergence of a sequence of vectors to convergence of the coordinate vector in \( \mathbb{F}^n \). It is a corollary of the above two fundamental lemmas.

**Definition 11.5.1** Let \((V, \|\cdot\|)\) be a normed linear space and let a basis be \( \{v_1, \cdots, v_n\} \). For \( x \in V \), let its component vector in \( \mathbb{F}^n \) be \( (\alpha_1, \cdots, \alpha_n) \) so that \( x = \sum \alpha_i v_i \). Then

\[
\|x\|_\infty = \max \{|\alpha_i|, i = 1, \cdots, n\}
\]

This is clearly a norm on \( V \). All axioms are obvious except for the triangle inequality. Say \( x, y \in V \) and let their coordinate vectors in \( \mathbb{F}^n \) be \( a, b \) respectively. Then

\[
\|x + y\|_\infty \equiv \max_i \{|a_i + b_i|\} \leq \max_i \{|a_i| + |b_i|\} \leq \max_i |a_i| + \max_i |b_i| \equiv \|x\|_\infty + \|y\|_\infty
\]

**Lemma 11.5.2** Let \((V, \|\cdot\|)\) be a finite dimensional vector space with basis equal to \( \{v_1, \cdots, v_n\} \) and for \( v \in V \), denote by \( v = (\alpha_1, \cdots, \alpha_n) \) the coordinate vector of \( v \) with respect to the given basis. Thus

\[
v = \sum_{k=1}^{n} \alpha_k v_k
\]

Then \( \|v^m\| \to 0 \) if and only if \( \|v^m\|_\infty \to 0 \).

**Proof:** \( \iff \) This direction is very easy. If the coordinate vectors converges to \( 0 \), \( (\|v^m\| \to 0) \) then

\[
\left\| \sum_{k=1}^{n} \alpha_k^m v_k \right\| \leq \sum_{k=1}^{n} |\alpha_k^m| \|v_k\| \leq \|v^m\|_\infty \sum_{k=1}^{n} \|v_k\|
\]

which converges to 0.

\( \Rightarrow \) Suppose not. Then there exists \( \varepsilon > 0 \) and a subsequence, still denoted by \( \{v^m\} \) such that \( \|v^m\|_\infty \geq \varepsilon \) although \( \|v^m\| \to 0 \). Then let \( w^m \equiv v^m / \|v^m\|_\infty \). It follows that

\[
\|w^m\| \leq \frac{1}{\varepsilon} \|v^m\|
\]
and so \( \|w^m\| \to 0 \) also. Letting

\[
w^m = \sum_{k=1}^{n} \beta_k^m v_k,\]

it follows that \( w^m = (\beta_1^m, \ldots, \beta_n^m) \) satisfies \( \|w^m\|_\infty = 1 \). Therefore, there is a further subsequence, still denoted with the index \( m \) and a single \( l \leq n \) such that \( |\beta_l^m| = 1 \) for all \( m \) and also each \( |\beta_k^m| \leq 1 \). It follows that for each \( k, \Re \beta_k^m \in [0, 1] \) and \( \Im \beta_k^m \in [0, 1] \). Therefore taking a succession of subsequences, there exists a further subsequence still denoted with the index \( m \) such that for every \( k \),

\[
\lim_{m \to \infty} \Re \beta_k^m = \gamma_k, \quad \lim_{m \to \infty} \Im \beta_k^m = \delta_k, \quad \beta_k = \gamma_k + i\delta_k
\]

As to the special case where \( k = l \),

\[
\lim_{m \to \infty} \sqrt{\Re(\beta_1^m)^2 + (\Im \beta_1^m)^2} = \lim_{m \to \infty} 1 = \sqrt{\gamma_1^2 + \delta_1^2}
\]

Then by the above observation in the first part of the proof,

\[
0 = \lim_{m \to \infty} \sum_{k=1}^{n} \beta_k^m v_k = \sum_{k=1}^{n} \beta_k v_k, \quad \beta_1 \neq 0.
\]

This contradicts the independence of the basis. Therefore, \( \|v^m\|_\infty \to 0 \) after all. 

Now here is the main theorem which states that in a finite dimensional normed linear space, any two norms are equivalent.

**Theorem 11.5.3** Let \( V \) be a finite dimensional vector space with basis \( \{v_1, \ldots, v_n\} \).

Also let \( \|\cdot\|_1, \|\cdot\|_2 \) be two norms. Then there exist positive constants \( \delta, \Delta \) such that for all \( v \in V \),

\[
\delta \|v\|_1 \leq \|v\|_2 \leq \Delta \|v\|_1.
\]

(11.13)

Also \( (V, \|\cdot\|) \) is automatically complete since \( V \) has finite dimensions.

**Proof:** We show that there exists a constant \( \Delta \) such that for all \( v \in V \), \( \|v\|_2 \leq \Delta \|v\|_1 \). If this is not so, then there exists a sequence \( v^m \) such that

\[
\|v^m\|_2 > n \|v^m\|_1
\]

Now let \( w^m = v^m / \|v^m\|_2 \). Then

\[
1 = \|w^m\|_2 > n \|w^m\|_1
\]

However, this requires \( \|w^m\|_1 \to 0 \) which implies \( \|w^m\|_\infty \to 0 \) which implies \( \|w^m\|_2 \to 0 \) contrary to \( \|w^m\|_2 = 1 \) for all \( m \). Hence there exists \( \Delta \) such that \( \|v\|_2 \leq \Delta \|v\|_1 \). Similarly, there exists a constant \( 1/\delta \) such that for all \( v \in V \), \( \|v\|_1 \leq \left(\frac{1}{\delta}\right) \|v\|_2 \). This proves the existence of the desired constants.

Consider the claim about completeness. The equivalence of any two norms in \( \mathbb{R^n} \) implies right away that completeness in one norm is equivalent to completeness in the other. Thus it suffices to show that \( (V, \|\cdot\|_\infty) \) is complete. Define a mapping \( \theta : (V, \|\cdot\|_\infty) \to \mathbb{R^n} \) as follows.

\[
\theta v \equiv \alpha
\]

where \( \alpha \) is the coordinate vector of \( v \) relative to the given basis. Thus \( \theta \) is one to one and onto and \( \|\theta v\|_\infty = \|v\|_\infty \) so both \( \theta \) and its inverse are continuous. If \( \{w_k\}_{k=1}^\infty \) is a Cauchy sequence in \( (V, \|\cdot\|_\infty) \), it follows that \( \{\theta w_k\} \) is a Cauchy sequence in \( \mathbb{R^n} \) which is complete by Example [Ex]. Therefore, there exists a unique \( \beta \) such that \( \theta w_k \to \beta \). Therefore, by continuity of \( \theta, w_k \to \theta^{-1} \beta \in V \) and so \( (V, \|\cdot\|_\infty) \) is complete. The equivalence of norms implies that for any norm on \( V; (V, \|\cdot\|) \) is complete.

It follows right away that the closed and open sets are the same with two different norms.
Corollary 11.5.4 Consider the metric spaces \((V, \| \cdot \|_1), (V, \| \cdot \|_2)\) where \(V\) has dimension \(n\). Then a set is closed or open in one of these if and only if it is respectively closed or open in the other. In other words, the two metric spaces have exactly the same open and closed sets. Also, a set is bounded in one metric space if and only if it is bounded in the other.

**Proof:** This follows from Theorem 11.5.3, the theorem about the equivalent formulations of continuity. Using this theorem, it follows from Theorem 11.5.3 that the identity map \(I (x) \equiv x\) is continuous. The reason for this is that the inequality of this theorem implies that if \(\|v^m - v\|_1 \to 0\) then \(\|Iv^m - Iv\|_2 = \|I (v^m - v)\|_2 \to 0\) and the same holds on switching 1 and 2 in what was just written.

Therefore, the identity map takes open sets to open sets and closed sets to closed sets. In other words, the two metric spaces have the same open sets and the same closed sets.

Suppose \(S\) is bounded in \((V, \| \cdot \|_1)\). This means it is contained in \(B (0, r)_1\) where the subscript of 1 indicates the norm is \(\| \cdot \|_1\). Let \(\delta \| \cdot \|_1 \leq \| \cdot \|_2 \leq \Delta \| \cdot \|_1\) as described above. Then 
\[
S \subseteq B (0, r)_1 \subseteq B (0, \Delta r)_2
\]
so \(S\) is also bounded in \((V, \| \cdot \|_2)\). Similarly, if \(S\) is bounded in \(\| \cdot \|_2\) then it is bounded in \(\| \cdot \|_1\).

### 11.6 Norms On \(\mathcal{L} (X, Y)\)

First here is a definition which applies in all cases, even if \(X, Y\) are infinite dimensional.

**Definition 11.6.1** Let \(X\) and \(Y\) be normed linear spaces with norms \(\| \cdot \|_X\) and \(\| \cdot \|_Y\) respectively. Then \(\mathcal{L} (X, Y)\) denotes the space of linear transformations, called bounded linear transformations, mapping \(X\) to \(Y\) which have the property that 
\[
\| A \| \equiv \sup \{ \| Ax \|_Y : \| x \|_X \leq 1 \} < \infty.
\]
Then \(\| A \|\) is referred to as the operator norm of the bounded linear transformation \(A\).

It is an easy exercise to verify that \(\| \cdot \|\) is a norm on \(\mathcal{L} (X, Y)\) and it is always the case that 
\[
\| Ax \|_Y \leq \| A \| \| x \|_X.
\]
Furthermore, you should verify that you can replace \(\leq 1\) with \(= 1\) in the definition. Thus 
\[
\| A \| \equiv \sup \{ \| Ax \|_Y : \| x \|_X = 1 \}.
\]

**Theorem 11.6.2** Let \(X\) and \(Y\) be finite dimensional normed linear spaces of dimension \(n\) and \(m\) respectively and denote by \(\| \cdot \|\) the norm on either \(X\) or \(Y\). Then if \(A\) is any linear function mapping \(X\) to \(Y\), then \(A \in \mathcal{L} (X, Y)\) and \((\mathcal{L} (X, Y), \| \cdot \|)\) is a complete normed linear space of dimension \(nm\) with 
\[
\| Ax \| \leq \| A \| \| x \|.
\]
Also if \(A \in \mathcal{L} (X, Y)\) and \(B \in \mathcal{L} (Y, Z)\) where \(X, Y, Z\) are normed linear spaces,
\[
\| BA \| \leq \| B \| \| A \|.
\]

**Proof:** It is necessary to show the norm defined on linear transformations really is a norm. Again the triangle inequality is the only property which is not obvious. It remains to show this and verify \(\| A \| < \infty\). Let \(\{v_1, \cdots, v_n\}\) be a basis and \(\| \cdot \|_\infty\) defined with respect to this basis as above, there exist constants \(\delta, \Delta > 0\) such that 
\[
\delta \| x \| \leq \| x \|_\infty \leq \Delta \| x \|.
\]
Then,
\[
\| A + B \| \equiv \sup \{ \| (A + B) (x) \| : \| x \| \leq 1 \}
\]
Next consider the claim that \( \|A\| < \infty \). This follows from

\[
\|A(x)\| = \|A\left(\sum_{i=1}^{n} x_{i}v_{i}\right)\| \leq \sum_{i=1}^{n} |x_{i}| \|A(v_{i})\|
\]

Thus \( \|A\| \leq \Delta \sum_{i=1}^{n} \|A(v_{i})\| \).

Next consider the assertion about the dimension of \( \mathcal{L}(X,Y) \). It follows from Theorem 5.1.4. By Theorem 11.5.3, \( \mathcal{L}(X,Y) \) is complete. If \( x \neq 0 \),

\[
\frac{\|Ax\|}{\|x\|} = \left\| A \frac{x}{\|x\|} \right\| \leq \|A\|
\]

Consider the last claim.

\[
\|BA\| = \sup_{\|x\| \leq 1} \|B(A(x))\| \leq \|B\| \sup_{\|x\| \leq 1} \|Ax\| = \|B\| \|A\| \]

Note by Theorem 11.5.3 you can define a norm any way desired on any finite dimensional linear space which has the field of scalars \( \mathbb{R} \) or \( \mathbb{C} \) and any other way of defining a norm on this space yields an equivalent norm. Thus, it doesn’t much matter as far as notions of convergence are concerned which norm is used for a finite dimensional space. In particular in the space of \( m \times n \) matrices, you can use the operator norm defined above, or some other way of giving this space a norm. A popular choice for a norm is the Frobenius norm.

**Definition 11.6.3** Make the space of \( m \times n \) matrices into a inner product space by defining

\[
(A,B) \equiv \text{trace}(AB^*) \equiv \sum_{i} (AB^*)_{ii} = \sum_{i} \sum_{j} A_{ij}B_{ji}^* = \sum_{i,j} A_{ij}B_{ij}
\]

\[\|A\| \equiv (A,A)^{1/2} \]

This is clearly a norm because, as implied by the notation, \( A,B \to (A,B) \) is an inner product on the space of \( m \times n \) matrices. You should verify that this is the case.

### 11.7 The Heine Borel Theorem

It has now been shown that the concept of closed or open set is totally independent of the particular norm used. Also, the issue of whether a set is bounded is independent of the choice of norm. The Heine Borel theorem says that the compact sets in \( \mathbb{F}^{n} \) where the distance is given by any norm are exactly those which are both closed and bounded. First here is a simple lemma.

**Lemma 11.7.1** Let \( (V,\|\cdot\|) \) be a normed linear space. Define

\[D(x,r) \equiv \{ \tilde{x} \in V : \|\tilde{x} - x\| \leq r \} \]

Then \( D(x,r) \) is a closed set. It is called a closed ball.

**Proof:** Let \( y \) be a limit point of \( D(x,r) \). Then by Theorem 11.5.3, there exists a sequence of distinct points of \( D(x,r) \), \( \{x_{n}\} \) none of which equal \( y \) such that \( x_{n} \to y \). Then

\[\|x_{n}\| - \|y\| \leq \|x_{n} - y\| \]

and so \( \|x_{n}\| \to \|y\| \). However, \( \|x_{n}\| \leq r \) and so it follows that \( \|y\| \leq r \) also. Hence this is a closed set as claimed because it contains all of its limit points. \( \blacksquare \)
11.7. THE HEINE BOREL THEOREM

Definition 11.7.2 A set $K$ in a metric space is compact if whenever $\mathcal{C}$ is a set of open balls whose union contains $K$, it follows that there is a finite subset of $\mathcal{C}$ whose union also contains $K$.

Now consider the case of $\mathbb{C}^n$ because it is slightly more technical.

Lemma 11.7.3 Now consider the case of $(\mathbb{F}^n, \|\cdot\|_\infty)$. In the notation of the above lemma, $D(0, r)_\infty = \{x \in \mathbb{F}^n: \|x\|_\infty \leq r\}$ is compact.

Proof: From the definition of the norm,

$$D(0, r)_\infty = \prod_{k=1}^n D(0, r) = D(0, r)^n$$

where here $D(0, r)$ is the closed disk in $\mathbb{F}$, either $[-r, r]$ in case $\mathbb{F} = \mathbb{R}$ or

$$\{x + iy : x^2 + y^2 \leq 1\}$$

in case $\mathbb{F} = \mathbb{C}$. Now consider $\{x^n\}_{n=1}^\infty$ a sequence in $D(0, r)^n$. It follows that for each $k \leq n$, $\{\text{Re } x_k^n\}_{m=1}^\infty$ and $\{\text{Im } x_k^n\}_{m=1}^\infty$ are both sequences in $[-r, r]$. Then taking a subsequence $2n$ times, there exists a subsequence, still denoted with $m$ as the index such that for each $k \leq n$, both $\{\text{Re } x_k\}_{m=1}^\infty$ and $\{\text{Im } x_k\}_{m=1}^\infty$ converge to points $\text{Re } x_k \in [-r, r]$ and $\text{Im } x_k \in [-r, r]$. It follows from the definition of the norm that $x^n \to x$ where $x = (x_1, \cdots, x_n)$. This has shown that $D(0, r)_\infty$ is sequentially compact. By Theorem 11.7.5 this is the same as saying that this set is compact. $\blacksquare$

Now here is the Heine Borel theorem.

Theorem 11.7.4 A subset $K$ of $(\mathbb{F}^n, \|\cdot\|_\infty)$ is compact if and only if it is closed and bounded.

Proof: $\implies$ Let $K$ be compact. If it is not bounded, then there is a sequence of points of $K, \{k^m\}_{m=1}^\infty$ such that $\|k^m\| \geq m$. It follows that it cannot have a convergent subsequence. Hence $K$ is not sequentially compact and consequently it is not compact. It follows that $K$ is bounded. If $K$ is not closed, then there exists a limit point $k$ which is not in $K$. By Theorem 11.7, there is a sequence of distinct points having no repeats and none equal to $k$ denoted as $\{k^m\}_{m=1}^\infty$ such that $k^m \to k$. Then for each $l$,

$$\{k\} \cup \bigcup_{k \in \mathbb{N}} \{k^k\} = H_k$$

is a closed set. (Why?) Also, it is clear that, since $k \notin K, U_k = H_k^C$ is a collection of open sets which covers $K$ but which admits no finite subcover. This is because $U_k$ fails to include $k^{k+1}$ and the sets $U_k$ are increasing in $k$.

$\impliedby$ Next suppose $K$ is closed and bounded and let $\mathcal{C}$ be an open cover of $K$. Then $K \subseteq D(0, r)_\infty$. It follows that $K^C \cup \mathcal{C}$ is an open cover of $D(0, r)_\infty$. This closed ball was shown to be compact in the above lemma. Therefore, there exists a finite subset of $\mathcal{C}, \{U_1, \cdots, U_p\}$ such that $\{U_1, \cdots, U_p, K^C\}$ is an open cover of $D(0, r)_\infty$. It follows that $\{U_1, \cdots, U_p\}$ is an open cover of $K$. Hence $K$ is compact.

This generalizes right away to an arbitrary finite dimensional normed linear space.

Theorem 11.7.5 (Heine Borel) Let $(V, \|\cdot\|)$ be a finite dimensional normed linear space with field of scalars $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$. Then the compact sets are exactly those which are closed and bounded. Also, any finite dimensional normed linear space is complete.

Proof: Let $\{v_1, \cdots, v_n\}$ be a basis. As before, for $v \in V$, let $\mathbf{v} = (\alpha_1, \cdots, \alpha_n)$ be the coordinate vector. Now define another norm $\|v\|_1 \equiv \|v\|_\infty$. It is left as an exercise to show that this is a norm on $V$. Then by Theorem 11.5.3, the one about equivalence of norms, there are positive constants $\delta, \Delta$ such that for all $v \in V$,

$$\delta \|v\|_\infty = \delta \|v\|_1 \leq \|v\| \leq \Delta \|v\|_1 = \Delta \|v\|_\infty.$$

Thus the closed and bounded sets in $V$ are the same for each of these two norms.
Let $K$ be closed and bounded in $(V, \|\cdot\|)$. Denote by $K^n$ the set of coordinate vectors from $K$. Then the above inequality shows that $K^n$ is closed and bounded in $\mathbb{F}^n$. If $\{v^m\}_{m=1}^{\infty}$ is a sequence in $K$, then $\{v^m\}_{m=1}^{\infty}$ is Cauchy in the compact set $K^n$ and so there is a subsequence, still denoted as $\{v^m\}_{m=1}^{\infty}$ which converges to $v \in K^n$. Letting $v$ be the vector of $K$ whose coordinate vector is $v$, it follows from Lemma 11.8.2 that $v^m \to v$. Thus if $K$ is closed and bounded, it is compact. As for the other direction, this follows right away from the first part of the argument given in the proof of Theorem 11.7.4 with no change.

Consider the last claim. Let $\{v^m\}_{m=1}^{\infty}$ be a Cauchy sequence in $(V, \|\cdot\|)$. Then it is contained in some closed and bounded set of the form $D(0, r)$. Consequently, it has a convergent subsequence. By Theorem 11.7.4 the original sequence converges also.

One can show that in the case of $\mathbb{R}$ where it makes sense to consider sup and inf, convergence of Cauchy sequences can be shown to imply the other definition of completeness involving sup, and inf.

### 11.8 Limits Of A Function

As in the case of scalar valued functions of one variable, a concept closely related to continuity is that of the limit of a function. The notion of limit of a function makes sense at points $x$, which are limit points of $D(f)$ and this concept is defined next. In all that follows $(V, \|\cdot\|)$ and $(W, \|\cdot\|)$ are two normed linear spaces. Recall the definition of limit point first.

**Definition 11.8.1** Let $A \subseteq W$ be a set. A point $x$, is a limit point of $A$ if $B(x, r)$ contains infinitely many points of $A$ for every $r > 0$.

**Definition 11.8.2** Let $f : D(f) \subseteq V \to W$ be a function and let $x$ be a limit point of $D(f)$. Then

$$\lim_{y \to x} f(y) = L$$

if and only if the following condition holds. For all $\varepsilon > 0$ there exists $\delta > 0$ such that if

$$0 < \|y - x\| < \delta, \text{ and } y \in D(f)$$

then,

$$\|L - f(y)\| < \varepsilon.$$

**Theorem 11.8.3** If $\lim_{y \to x} f(y) = L$ and $\lim_{y \to x} f(y) = L_1$, then $L = L_1$.

**Proof:** Let $\varepsilon > 0$ be given. There exists $\delta > 0$ such that if $0 < \|y - x\| < \delta$ and $y \in D(f)$, then

$$\|f(y) - L\| < \varepsilon, \|f(y) - L_1\| < \varepsilon.$$

Pick such a $y$. There exists one because $x$ is a limit point of $D(f)$. Then

$$\|L - L_1\| \leq \|L - f(y)\| + \|f(y) - L_1\| < \varepsilon + \varepsilon = 2\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this shows $L = L_1$. $\blacksquare$

As in the case of functions of one variable, one can define what it means for $\lim_{y \to x} f(x) = \pm \infty$.

**Definition 11.8.4** If $f(x) \in \mathbb{R}$, $\lim_{y \to x} f(x) = \infty$ if for every number $l$, there exists $\delta > 0$ such that whenever $\|y - x\| < \delta$ and $y \in D(f)$, then $f(y) > l$. $\lim_{y \to x} f(x) = -\infty$ if for every number $l$, there exists $\delta > 0$ such that whenever $\|y - x\| < \delta$ and $y \in D(f)$, then $f(y) < l$.

The following theorem is just like the one variable version of calculus.
Theorem 11.8.5 Suppose $f : D (f) \subseteq V \to \mathbb{F}^m$. Then for $x$ a limit point of $D (f)$,  
\[ \lim_{y \to x} f (y) = L \]  
(11.14)  
if and only if  
\[ \lim_{y \to x} f_k (y) = L_k \]  
(11.15)  
where $f (y) \equiv (f_1 (y), \ldots, f_p (y))$ and $L \equiv (L_1, \ldots, L_p)$.

Suppose here that $f$ has values in $W$, a normed linear space and  
\[ \lim_{y \to x} f (y) = L, \quad \lim_{y \to x} g (y) = K \]  
where $K, L \in W$. Then if $a, b \in \mathbb{F}$,  
\[ \lim_{y \to x} (af (y) + bg (y)) = aL + bK, \]  
(11.16)  
If $W$ is an inner product space,  
\[ \lim_{y \to x} (f, g) (y) = (L, K) \]  
(11.17)  
If $g$ is scalar valued with $\lim_{y \to x} g (y) = K$,  
\[ \lim_{y \to x} f (y) g (y) = LK. \]  
(11.18)

Also, if $h$ is a continuous function defined near $L$, then  
\[ \lim_{y \to x} h \circ f (y) = h (L). \]  
(11.19)  
Suppose $\lim_{y \to x} f (y) = L$. If $\| f (y) - b \| \leq r$ for all $y$ sufficiently close to $x$, then $|L - b| \leq r$ also.

Proof: Suppose $\mathbb{R}^m$. Then letting $\varepsilon > 0$ be given there exists $\delta > 0$ such that if $0 < \| y - x \| < \delta$, it follows  
\[ |f_k (y) - L_k| \leq \| f (y) - L \| < \varepsilon \]  
which verifies $\mathbb{R}^m$.

Now suppose $\mathbb{R}^m$ holds. Then letting $\varepsilon > 0$ be given, there exists $\delta_k$ such that if $0 < \| y - x \| < \delta_k$, then  
\[ |f_k (y) - L_k| < \varepsilon. \]  
Let $0 < \delta < \min (\delta_1, \ldots, \delta_p)$. Then if $0 < \| y - x \| < \delta$, it follows  
\[ \| f (y) - L \|_\infty < \varepsilon \]  
Any other norm on $\mathbb{F}^m$ would work out the same way because the norms are all equivalent.

Each of the remaining assertions follows immediately from the coordinate descriptions of the various expressions and the first part. However, I will give a different argument for these.

The proof of $\mathbb{R}^m$ is left for you. Now $\mathbb{R}^m$ is to be verified. Let $\varepsilon > 0$ be given. Then by the triangle inequality,  
\[ |(f, g) (y) - (L, K)| \leq |(f, g) (y) - (f (y), K)| + |(f (y), K) - (L, K)| \leq \| f (y) \| \| g (y) - K \| + \| K \| \| f (y) - L \|. \]  
There exists $\delta_1$ such that if $0 < \| y - x \| < \delta_1$ and $y \in D (f)$, then  
\[ \| f (y) - L \| < 1, \]  
and so for such $y$, the triangle inequality implies, $\| f (y) \| < 1 + \| L \|$. Therefore, for $0 < \| y - x \| < \delta_1$,  
\[ |(f, g) (y) - (L, K)| \leq (1 + \| K \| + \| L \|) (\| g (y) - K \| + \| f (y) - L \|). \]  
(11.20)
Now let \(0 < \delta_2\) be such that if \(y \in D(f)\) and \(0 < \|x-y\| < \delta_2\),
\[
   \|f(y) - L\| < \frac{\varepsilon}{2(1+\|K\| + \|L\|)}, \quad \|g(y) - K\| < \frac{\varepsilon}{2(1+\|K\| + \|L\|)}.
\]

Then letting \(\delta < \min(\delta_1, \delta_2)\), it follows from \(\|\cdot\|\) that
\[
   |(f,g)(y) - (L,K)| < \varepsilon
\]
and this proves \(\|\cdot\|\).

The proof of \(\|\cdot\|\) is left to you.

Consider \(\|\cdot\|\). Since \(h\) is continuous near \(L\), it follows that for \(\varepsilon > 0\) given, there exists \(\eta > 0\) such that if \(\|y-L\| < \eta\), then
\[
   \|h(y) - h(L)\| < \varepsilon
\]
Now since \(\lim_{y \to x} f(y) = L\), there exists \(\delta > 0\) such that if \(0 < \|y-x\| < \delta\), then
\[
   \|f(y) - L\| < \eta.
\]
Therefore, if \(0 < \|y-x\| < \delta\),
\[
   \|h(f(y)) - h(L)\| < \varepsilon.
\]

It only remains to verify the last assertion. Assume \(\|f(y) - b\| \leq r\). It is required to show that \(\|L-b\| \leq r\). If this is not true, then \(\|L-b\| > r\). Consider \(B(L, \|L-b\| - r)\). Since \(L\) is the limit of \(f\), it follows \(f(y) \in B(L, \|L-b\| - r)\) whenever \(y \in D(f)\) is close enough to \(x\). Thus, by the triangle inequality,
\[
   \|f(y) - L\| < \|L-b\| - r
\]
and so
\[
   r < \|L-b\| - \|f(y) - L\| \leq \|b-L\| - \|f(y) - L\|
\]
\[
   \leq \|b-f(y)\|,
\]
a contradiction to the assumption that \(\|b-f(y)\| \leq r\).

The relation between continuity and limits is as follows.

**Theorem 11.8.6** For \(f : D(f) \to W\) and \(x \in D(f)\) a limit point of \(D(f)\), \(f\) is continuous at \(x\) if and only if
\[
   \lim_{y \to x} f(y) = f(x).
\]

**Proof:** First suppose \(f\) is continuous at \(x\) a limit point of \(D(f)\). Then for every \(\varepsilon > 0\) there exists \(\delta > 0\) such that if \(\|x-y\| < \delta\) and \(y \in D(f)\), then \(|f(x) - f(y)| < \varepsilon\). In particular, this holds if \(0 < \|x-y\| < \delta\) and this is just the definition of the limit. Hence \(f(x) = \lim_{y \to x} f(y)\).

Next suppose \(x\) is a limit point of \(D(f)\) and \(\lim_{y \to x} f(y) = f(x)\). This means that if \(\varepsilon > 0\) there exists \(\delta > 0\) such that for \(0 < \|x-y\| < \delta\) and \(y \in D(f)\), it follows \(|f(y) - f(x)| < \varepsilon\). However, if \(y = x\), then \(|f(y) - f(x)| = |f(x) - f(x)| = 0\) and so whenever \(y \in D(f)\) and \(\|x-y\| < \delta\), it follows \(|f(x) - f(y)| < \varepsilon\), showing \(f\) is continuous at \(x\).

**Example 11.8.7** Find \(\lim_{(x,y) \to (3,1)} \left(\frac{x^2 - 9}{x-3}, y\right)\).

It is clear that \(\lim_{(x,y) \to (3,1)} \frac{x^2 - 9}{x-3} = 6\) and \(\lim_{(x,y) \to (3,1)} y = 1\). Therefore, this limit equals \((6, 1)\).

**Example 11.8.8** Find \(\lim_{(x,y) \to (0,0)} \frac{xy}{x^2 + y^2}\).


11.9. **Exercises**

1. Consider the metric space $C([0, T], \mathbb{R}^n)$ with the norm $\|f\| \equiv \max_{x \in [0, T]} \|f(x)\|_\infty$. Explain why the maximum exists. Show this is a complete metric space. **Hint:** If you have $\{f_m\}$ a Cauchy sequence in $C([0, T], \mathbb{R}^n)$, then for each $x$, you have $\{f_m(x)\}$ a Cauchy sequence in $\mathbb{R}^n$ so it converges by completeness of $\mathbb{R}^n$. See Example 11.9.2. Thus there exists $f(x) \equiv \lim_{m \to \infty} f_m(x)$. You must show that $f$ is continuous. Consider

$$
\|f_m(x) - f_m(y)\| \leq \|f_m(x) - f_n(x)\| + \|f_n(x) - f_n(y)\| + \|f_n(y) - f_m(y)\|
$$

for $n$ large enough. Now let $m \to \infty$ to get the same inequality with $f$ on the left. Next use continuity of $f_n$. Finally,

$$
\|f(x) - f_n(x)\| = \lim_{m \to \infty} \|f_m(x) - f_n(x)\|
$$

and since a Cauchy sequence, $\|f_m - f_n\| < \varepsilon$ whenever $m > n$ for $n$ large enough. Use to show that $\|f - f_n\|_\infty \to 0$.

2. For $f \in C([0, T], \mathbb{R}^n)$, you define the Riemann integral in the usual way using Riemann sums. Alternatively, you can define it as

$$
\int_0^t f(s) \, ds = \left( \int_0^t f_1(s) \, ds, \int_0^t f_2(s) \, ds, \ldots, \int_0^t f_n(s) \, ds \right)
$$

Then show that the following limit exists in $\mathbb{R}^n$ for each $t \in (0, T)$.

$$
\lim_{h \to 0} \frac{\int_0^{t+h} f(s) \, ds - \int_0^t f(s) \, ds}{h} = f(t)
$$

You should use the fundamental theorem of calculus from one variable calculus and the definition of the norm to verify this. Recall that

$$
\lim_{t \to \infty} f(t) = 1
$$

means that for all $\varepsilon > 0$, there exists $\delta > 0$ such that if $0 < |t - s| < \delta$, then $\|f(t) - 1\|_\infty < \varepsilon$. You have to use the definition of a limit in order to establish that something is a limit.

3. A collection of functions $F$ of $C([0, T], \mathbb{R}^n)$ is said to be uniformly equicontinuous if for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $f \in F$ and $|t - s| < \delta$, then $\|f(t) - f(s)\|_\infty < \varepsilon$. Thus the functions are uniformly continuous all at once. The single $\delta$ works for every pair $t, s$ closer together than $\delta$ and for all functions $f \in F$. As an easy case, suppose there exists $K$ such that for all $f \in F$,

$$
\|f(t) - f(s)\|_\infty \leq K |t - s|
$$
show that $\mathcal{F}$ is uniformly equicontinuous. Now suppose $\mathcal{G}$ is a collection of functions of $C ([0, T], \mathbb{R}^n)$ which is bounded. That is, $\|f\| = \max_{t \in [0, T]} \|f(t)\|_{\infty} < M < \infty$ for all $f \in \mathcal{G}$. Then let $\mathcal{F}$ denote the functions which are of the form

$$F(t) \equiv y_0 + \int_0^t f(s) \, ds$$

where $f \in \mathcal{G}$. Show that $\mathcal{F}$ is uniformly equicontinuous. **Hint:** This is a really easy problem if you do the right things. Here is the way you should proceed. Remember the triangle inequality from one variable calculus which said that for $a < b$ \[ \int_a^b |f(s)| \, ds \leq \int_a^b |f(s)| \, ds \leq \int_a^b \|f\|_{\infty} \, ds \]

Reduce to the case just considered using the assumption that these $f$ are bounded.

4. Let $V$ be a vector space with basis $\{v_1, \ldots, v_n\}$. For $v \in V$, denote its coordinate vector as $v = (\alpha_1, \ldots, \alpha_n)$ where $v = \sum_{k=1}^n \alpha_k v_k$. Now define

$$\|v\| \equiv \|v\|_{\infty}$$

Show that this is a norm on $V$.

5. Let $(X, \|\|)$ be a normed linear space. A set $A$ is said to be convex if whenever $x, y \in A$ the line segment determined by these points given by $tx + (1 - t)y$ for $t \in [0, 1]$ is also in $A$. Show that every open or closed ball is convex. Remember a closed ball is $D(x, r) \equiv \{\xi : \|\xi - x\| \leq r\}$ while the open ball is $B(x, r) \equiv \{\xi : \|\xi - x\| < r\}$. This should work just as easily in any normed linear space with any norm.

6. A vector $v$ is in the convex hull of a nonempty set $S$ if there are finitely many vectors of $S, \{v_1, \ldots, v_m\}$ and nonnegative scalars $\{t_1, \ldots, t_m\}$ such that

$$v = \sum_{k=1}^m t_k v_k, \quad \sum_{k=1}^m t_k = 1.$$ 

Such a linear combination is called a convex combination. Suppose now that $S \subseteq V$, a vector space of dimension $n$. Show that if $v = \sum_{k=1}^m t_k v_k$ is a vector in the convex hull for $m > n + 1$, then there exist other nonnegative scalars $\{t'_k\}$ summing to 1 such that

$$v = \sum_{k=1}^{m-1} t'_k v_k.$$ 

Thus every vector in the convex hull of $S$ can be obtained as a convex combination of at most $n + 1$ points of $S$. This incredible result is in Rudin [9]. Convexity is more a geometric property than a topological property. **Hint:** Consider $L : \mathbb{R}^m \to V \times \mathbb{R}$ defined by

$$L(a) \equiv \left( \sum_{k=1}^m a_k v_k, \sum_{k=1}^m a_k \right)$$

Explain why $\ker(L) \neq \{0\}$. This will involve observing that $\mathbb{R}^m$ has higher dimension that $V \times \mathbb{R}$. Thus $L$ cannot be one to one because one to one functions take linearly independent sets to linearly independent sets and you can’t have a linearly independent set with more than $n + 1$ vectors in $V \times \mathbb{R}$. Next, letting $a \in \ker(L) \setminus \{0\}$ and $\lambda \in \mathbb{R}$, note that $\lambda a \in \ker(L)$. Thus for all $\lambda \in \mathbb{R}$,

$$v = \sum_{k=1}^m (t_k + \lambda a_k) v_k.$$
Now vary $\lambda$ till some $t_k + \lambda a_k = 0$ for some $a_k \neq 0$. You can assume each $t_k > 0$ since otherwise, there is nothing to show. This is a really nice result because it can be used to show that the convex hull of a compact set is also compact. Show this next. This is also Problem 22 but here it is again. This is because it is a really nice result.

7. Show that the usual norm in $\mathbb{F}^n$ given by

$$|x| = (x, x)^{1/2}$$

satisfies the following identities, the first of them being the parallelogram identity and the second being the polarization identity.

$$|x + y|^2 + |x - y|^2 = 2|x|^2 + 2|y|^2$$

$$\text{Re}(x, y) = \frac{1}{4} \left(|x + y|^2 - |x - y|^2\right)$$

Show that these identities hold in any inner product space, not just $\mathbb{F}^n$.

8. Let $K$ be a nonempty closed and convex set in an inner product space $(X, |.|)$ which is complete. For example, $\mathbb{F}^n$ or any other finite dimensional inner product space. Let $y \notin K$ and let

$$\lambda = \inf \{ |y - x| : x \in K \}$$

Let $\{x_n\}$ be a minimizing sequence. That is

$$\lambda = \lim_{n \to \infty} |y - x_n|$$

Explain why such a minimizing sequence exists. Next explain the following using the parallelogram identity in the above problem as follows.

$$\left| y - \frac{x_n + x_m}{2} \right|^2 = \frac{|y - x_n|^2 + |y + x_m|^2}{2}$$

$$= -\left| y - \frac{x_n}{2} - \left(\frac{y}{2} - \frac{x_m}{2}\right) \right|^2 + \frac{1}{2} |y - x_n|^2 + \frac{1}{2} |y - x_m|^2$$

Hence

$$\left| \frac{x_m - x_n}{2} \right|^2 = -\left| y - \frac{x_n + x_m}{2} \right|^2 + \frac{1}{2} |y - x_n|^2 + \frac{1}{2} |y - x_m|^2$$

$$\leq -\lambda^2 + \frac{1}{2} |y - x_n|^2 + \frac{1}{2} |y - x_m|^2$$

Next explain why the right hand side converges to 0 as $m, n \to \infty$. Thus $\{x_n\}$ is a Cauchy sequence and converges to some $x \in X$. Explain why $x \in K$ and $|x - y| = \lambda$. Thus there exists a closest point in $K$ to $y$. Next show that there is only one closest point. Hint: To do this, suppose there are two $x_1, x_2$ and consider $\frac{x_1 + x_2}{2}$ using the parallelogram law to show that this average works better than either of the two points which is a contradiction unless they are really the same point. This theorem is of enormous significance.

9. Let $K$ be a closed convex nonempty set in a complete inner product space $(H, |.|)$ (Hilbert space) and let $y \in H$. Denote the closest point to $y$ by $Py$. Show that $Py$ is characterized as being the solution to the following variational inequality

$$\text{Re}(z - Py, y - Py) \leq 0$$

for all $z \in K$. That is, show that $x = Py$ if and only if $\text{Re}(z - x, y - x) \leq 0$ for all $z \in K$. Hint: Let $x \in K$. Then, due to convexity, a generic thing in $K$ is of the form $x + t(z - x), t \in [0, 1]$ for every $z \in K$. Then

$$|x + t(z - x) - y|^2 = |x - y|^2 + t^2 |z - x|^2 - t2 \text{Re}(z - x, y - x)$$
If \( x = Px \), then the minimum value of this on the left occurs when \( t = 0 \). Function defined on \([0,1]\) has its minimum at \( t = 0 \). What does it say about the derivative of this function at \( t = 0 \)? Next consider the case that for some \( x \) the inequality \( \text{Re} (z - x, y - x) \leq 0 \). Explain why this shows \( x = Py \).

10. Using Problem 3 and Problem 8 show the projection map, \( P \) onto a closed convex subset is Lipschitz continuous with Lipschitz constant 1. That is

\[
|Py - Py_0| \leq |x - y|
\]

11. Suppose, in an inner product space, you know \( \text{Re} (x, y) \). Show that you also know \( \text{Im} (x, y) \). That is, give a formula for \( \text{Im} (x, y) \) in terms of \( \text{Re} (x, y) \). Hint:

\[
(x, iy) = -i (x, y) = -i (\text{Re} (x, y) + i \text{Im} (x, y))
\]

while, by definition,

\[
(x, iy) = \text{Re} (x, iy) + i \text{Im} (x, iy)
\]

Now consider matching real and imaginary parts.

12. Suppose \( K \) is a compact subset (If \( C \) is a set of open sets whose union contains \( K \), (open cover) then there are finitely many sets of \( C \) whose union contains \( K \) ) of \( (X, d) \) a metric space. Also let \( C \) be an open cover of \( K \). Show that there exists \( \delta > 0 \) such that for all \( x \in K \), \( B (x, \delta) \) is contained in a single set of \( C \). This number is called a Lebesgue number. Hint: For each \( x \in K \), there exists \( B (x, \delta_x) \) such that this ball is contained in a set of \( C \). Now consider the balls \( \{ B (x, \delta_x/2) \}_{x \in K} \). Finitely many of these cover \( K \). \( \{ B (x_i, \delta_{x_i}/2) \}_{i=1}^n \) Now consider what happens if you let \( \delta \leq \min \{ \delta_{x_i}/2, i = 1, 2, \ldots, n \} \). Explain why this works. You might draw a picture to help get the idea.

13. Suppose \( C \) is a set of compact sets in a metric space \( (X, d) \) and suppose that the intersection of every finite subset of \( C \) is nonempty. This is called the finite intersection property. Show that \( \cap C \), the intersection of all sets of \( C \) is nonempty. This particular result is enormously important. Hint: You could let \( U \) denote the set \( \{ K^C: K \in C \} \). If \( \cap C \) is empty, then its complement is \( \cup U = X \). Picking \( K \in C \), it follows that \( U \) is an open cover of \( K \). Therefore, you would need to have \( \{ K_1^C, \ldots, K_m^C \} \) is a cover of \( K \). In other words,

\[
K \subseteq \cup_{i=1}^m K_i^C = (\cap_{i=1}^m K_i)^C
\]

Now what does this say about the intersection of \( K \) with these \( K_i^C \)?

14. Show that if \( f \) is continuous and defined on a compact set \( K \) in a metric space, then it is uniformly continuous. Continuous means continuous at every point. Uniformly continuous means: For every \( \varepsilon > 0 \) there exists \( \delta > 0 \) such that if \( d(x, y) < \delta \), then \( d(f(x), f(y)) < \varepsilon \). The difference is that \( \delta \) does not depend on \( x \). Hint: Use the existence of the Lebesgue number in Problem 12 to prove uniformity on a compact set \( K \) implies uniform continuity on this set. Hint: Consider \( C \equiv \{ f^{-1} (B (f(x), \varepsilon/2)) : x \in X \} \). This is an open cover of \( X \). Let \( \delta \) be a Lebesgue number for this open cover. Suppose \( d(x, \tilde{x}) < \delta \). Then both \( x, \tilde{x} \) are in \( B (x, \delta) \) and so both are in \( f^{-1} (B (f(x), \varepsilon/2)) \). Hence \( \rho (f(x), f(\tilde{x})) < \varepsilon/2 \) and \( \rho (f(\tilde{x}), f(\tilde{x})) < \varepsilon/2 \) and now consider the triangle inequality. Recall the usual definition of continuity. In metric space it is as follows: For \( (D, d), (Y, \rho) \) metric spaces, \( f: D \to Y \) is continuous at \( x \in D \) means that for all \( \varepsilon > 0 \) there exists \( \delta > 0 \) such that if \( d(y, x) < \delta \), then \( \rho (f(x), f(y)) < \varepsilon \). Continuity on \( D \) means continuity at every point of \( D \).

15. The definition of compactness is that a set \( K \) is compact if and only if every open cover (collection of open sets whose union contains \( K \)) has a finite subset which is also an open cover. Show that this is equivalent to saying that every open cover consisting of balls has a finite subset which is also an open cover.
16. A set $K$ in a metric space is said to be sequentially compact if whenever \( \{x_n\} \) is a sequence in $K$, there exists a subsequence which converges to a point of $K$. Show that if $K$ is compact, then it is sequentially compact. **Hint:** Explain why if $x \in K$, then there exist an open set $B_x$ containing $x$ which has $x_k$ for only finitely many values of $k$. Then use compactness. This was shown in the chapter, but do your own proof of this part of it.

17. Show that $f: D \rightarrow Y$ is continuous at $x \in D$ where $(D,d)$, $(Y,\rho)$ are a metric spaces if and only if whenever $x_n \rightarrow x$ in $D$, it follows that $f(x_n) \rightarrow f(x)$. Recall the usual definition of continuity. $f$ is continuous at $x$ means that for all $\varepsilon > 0$ there exists $\delta > 0$ such that if $d(y,x) < \delta$, then $\rho(f(x), f(y)) < \varepsilon$. Continuity on $D$ means continuity at every point of $D$.

18. Give an easier proof of the result of Problem 14. **Hint:** If $f$ is not uniformly continuous, then there exists $\varepsilon > 0$ and $x_n, y_n, d(x_n, y_n) < \frac{1}{n}$ but $d(f(x_n), f(y_n)) \geq \varepsilon$. Now use sequential compactness of $K$ to get a contradiction.

19. This problem will reveal the best kept secret in undergraduate mathematics, the definition of the derivative of a function of $n$ variables. Let $\|\cdot\|_V$ be a norm on $V$ and also denote by $\|\cdot\|_W$ a norm on $W$. Write $\|\cdot\|$ for both to save notation. Let $U \subseteq V$ be an open set. Let $f: U \rightarrow W$ be a function having values in $W$. Then $f$ is differentiable at $x \in U$ means that there exists $A \in \mathcal{L}(V, W)$ such that for every $\varepsilon > 0$, there exists a $\delta > 0$ such that whenever $0 < \|v\| < \delta$, it follows that

$$\frac{\|f(x + v) - f(x) - Av\|}{\|v\|} < \varepsilon$$

Stated more simply,

$$\lim_{\|v\| \rightarrow 0} \frac{\|f(x + v) - f(x) - Av\|}{\|v\|} = 0$$

Show that $A$ is unique. It is written as $Df(x) = A$. This is what is meant by the derivative of $f$. If $V = \mathbb{R}^n$, and $W = \mathbb{R}^m$, show that with respect to the usual bases, the matrix of $Df(x)$ is an $m \times n$ matrix whose $k^{th}$ column is $\frac{\partial f}{\partial x_k}$.

20. Let $V, W$ be finite dimensional normed linear spaces and let $\{A_n\}_{n=1}^\infty$ be a sequence of linear transformations in $\mathcal{L}(V, W)$ such that $\sup_n \|A_n\| < \infty$. Show that there exists a subsequence $\{A_{n_k}\}$ and $A \in \mathcal{L}(V, W)$ such that $\lim_{n \rightarrow \infty} \|A_n - A\| = 0$.

21. Given an example of a sequence $\{A_k\} \subseteq \mathcal{L}(V, V)$ such that the minimum polynomial of each $A_k$ has degree $n = \dim(V)$ but $\|A_k - A\| \rightarrow 0$ and the minimum polynomial of $A$ has degree less than $n$. **Hint:** You might want to think in terms of the Jordan form.
Chapter 12

Limits Of Vectors And Matrices

12.1 Regular Markov Matrices

The existence of the Jordan form is the basis for the proof of limit theorems for certain kinds of matrices called Markov matrices.

Definition 12.1.1 An \( n \times n \) matrix \( A = (a_{ij}) \), is a Markov matrix if \( a_{ij} \geq 0 \) for all \( i, j \) and

\[
\sum_i a_{ij} = 1.
\]

It may also be called a stochastic matrix or a transition matrix. A Markov or stochastic matrix is called regular if some power of \( A \) has all entries strictly positive. A vector \( v \in \mathbb{R}^n \), is a steady state if \( Av = v \).

Lemma 12.1.2 The property of being a stochastic matrix is preserved by taking products. It is also true if the sum is of the form \( \sum_j a_{ij} = 1 \).

Proof: Suppose the sum over a row equals 1 for \( A \) and \( B \). Then letting the entries be denoted by \((a_{ij})\) and \((b_{ij})\) respectively and the entries of \( AB \) by \((c_{ij})\),

\[
\sum_i c_{ij} = \sum_i \sum_k a_{ik}b_{kj} = \sum_k \sum_i a_{ik}b_{kj} = \sum_k b_{kj} = 1
\]

It is obvious that when the product is taken, if each \( a_{ij}, b_{ij} \geq 0 \), then the same will be true of sums of products of these numbers. Similar reasoning works for the assumption that \( \sum_j a_{ij} = 1 \).

The following theorem is convenient for showing the existence of limits.

Theorem 12.1.3 Let \( A \) be a real \( p \times p \) matrix having the properties

1. \( a_{ij} \geq 0 \)
2. Either \( \sum_{i=1}^{p} a_{ij} = 1 \) or \( \sum_{j=1}^{p} a_{ij} = 1 \).
3. The distinct eigenvalues of \( A \) are \( \{\lambda_1, \ldots, \lambda_m\} \) where each \( |\lambda_j| < 1 \).

Then \( \lim_{n \to \infty} A^n = A_\infty \) exists in the sense that \( \lim_{n \to \infty} a_{ij}^n = a_{ij}^\infty \), the \( ij \)th entry \( A_\infty \). Here \( a_{ij}^n \) denotes the \( ij \)th entry of \( A^n \). Also, if \( \lambda = 1 \) has algebraic multiplicity \( r \), then the Jordan block corresponding to \( \lambda = 1 \) is just the \( r \times r \) identity.
Proof. By the existence of the Jordan form for $A$, it follows that there exists an invertible matrix $P$ such that

$$P^{-1}AP = \begin{pmatrix} I + N & J_{r_2}(\lambda_2) & \cdots & J_{r_m}(\lambda_m) \\ & & & \\
\end{pmatrix} = J$$

where $I$ is $r \times r$ for $r$ the multiplicity of the eigenvalue 1 and $N$ is a nilpotent matrix for which $N^r = 0$. I will show that because of Condition 2, $N = 0$.

First of all,

$$J_{r_i}(\lambda_i) = \lambda_i I + N_i$$

where $N_i$ satisfies $N_i^{r_i} = 0$ for some $r_i > 0$. It is clear that $N_i (\lambda_i I) = (\lambda_i I) N_i$ and so

$$(J_{r_i}(\lambda_i))^n = \sum_{k=0}^{n} \binom{n}{k} N_i^k \lambda_i^{n-k} = \sum_{k=0}^{r} \binom{n}{k} N_i^k \lambda_i^{n-k}$$

which converges to 0 due to the assumption that $|\lambda_i| < 1$. There are finitely many terms and a typical one is a matrix whose entries are no larger than an expression of the form

$$|\lambda_i|^{n-k} C_k n (n-1) \cdots (n-k+1) \leq C_k |\lambda_i|^{n-k} n^k$$

which converges to 0 because, by the root test, the series $\sum_{n=1}^{\infty} |\lambda_i|^{n-k} n^k$ converges. Thus for each $i = 2, \ldots, p$,

$$\lim_{n \to \infty} (J_{r_i}(\lambda_i))^n = 0.$$

By Condition 2, if $a_{ij}^n$ denotes the $ij^{th}$ entry of $A^n$, then either

$$\sum_{i=1}^{p} a_{ij}^n = 1 \text{ or } \sum_{j=1}^{p} a_{ij}^n = 1, \text{ } a_{ij}^n \geq 0.$$

This follows from Lemma 12.1.2. It is obvious each $a_{ij}^n \geq 0$, and so the entries of $A^n$ must be bounded independent of $n$.

It follows easily from

$$\underbrace{P^{-1}APP^{-1}AP \cdots P^{-1}AP}_{n \text{ times}} = P^{-1}A^n P$$

that

$$P^{-1}A^n P = J^n$$

Hence $J^n$ must also have bounded entries as $n \to \infty$. However, this requirement is incompatible with an assumption that $N \neq 0$.

If $N \neq 0$, then $N^s \neq 0$ but $N^{s+1} = 0$ for some $1 \leq s \leq r$. Then

$$(I + N)^n = I + \sum_{k=1}^{s} \binom{n}{k} N^k$$

One of the entries of $N^s$ is nonzero by the definition of $s$. Let this entry be $n_{ij}^s$. Then this implies that one of the entries of $(I + N)^n$ is of the form $\binom{n}{s} n_{ij}^s$. This entry dominates the $ij^{th}$ entries of $\binom{n}{k} N^k$ for all $k < s$ because

$$\lim_{n \to \infty} \frac{\binom{n}{s}}{\binom{n}{k}} = \infty$$
12.1. REGULAR MARKOV MATRICES

Therefore, the entries of \((I + N)^n\) cannot all be bounded. From block multiplication,

\[
P^{-1}A^n P = \begin{pmatrix}
(I + N)^n & (J_{r2}(\lambda_2))^n \\
. & . \\
. & (J_{rn}(\lambda_m))^n
\end{pmatrix}
\]

and this is a contradiction because entries are bounded on the left and unbounded on the right.

Since \(N = 0\), the above equation implies \(\lim_{n \to \infty} A^n\) exists and equals

\[
P \begin{pmatrix}
I \\
0 \\
. \\
0
\end{pmatrix}^{P^{-1} \mathbb{I}}
\]

Are there examples which will cause the eigenvalue condition of this theorem to hold? The following lemma gives such a condition. It turns out that if \(a_{ij} > 0\), not just \(\geq 0\), then the eigenvalue condition of the above theorem is valid.

**Lemma 12.1.4** Suppose \(A = (a_{ij})\) is a stochastic matrix. Then \(\lambda = 1\) is an eigenvalue. If \(a_{ij} > 0\) for all \(i, j\), then if \(\mu\) is an eigenvalue of \(A\), either \(|\mu| < 1\) or \(\mu = 1\).

**Proof:** First consider the claim that 1 is an eigenvalue. By definition,

\[
\sum_i 1a_{ij} = 1
\]

and so \(A^T \mathbf{v} = \mathbf{v}\) where \(\mathbf{v} = \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}^T\). Since \(A, A^T\) have the same eigenvalues, this shows 1 is an eigenvalue. Suppose then that \(\mu\) is an eigenvalue. Is \(|\mu| < 1\) or \(\mu = 1\)? Let \(\mathbf{v}\) be an eigenvector for \(A^T\) and let \(|v_i|\) be the largest of the \(|v_j|\).

\[
\mu v_i = \sum_j a_{ji}v_j
\]

and now multiply both sides by \(\overline{\mu} v_i\) to obtain

\[
|\mu|^2 |v_i|^2 = \sum_j a_{ji}v_j\overline{\mu} v_i = \sum_j a_{ji} \text{Re}(v_j\overline{\mu} v_i)
\]

\[
\leq \sum_j a_{ji} |v_i|^2 |\mu| = |\mu| |v_i|^2
\]

Therefore, \(|\mu| \leq 1\). If \(|\mu| = 1\), then equality must hold in the above, and so \(v_j\overline{\mu} v_i\) must be real and nonnegative for each \(j\). In particular, this holds for \(j = i\) which shows \(\overline{\mu}\) is real and nonnegative. Thus, in this case, \(\mu = 1\) because \(\overline{\mu} = \mu\) is nonnegative and equal to 1. The only other case is where \(|\mu| < 1\).

**Lemma 12.1.5** Let \(A\) be any Markov matrix and let \(\mathbf{v}\) be a vector having all its components non negative with \(\sum_i v_i = c\). Then if \(\mathbf{w} = A\mathbf{v}\), it follows that \(w_i \geq 0\) for all \(i\) and \(\sum_i w_i = c\).

**Proof:** From the definition of \(\mathbf{w}\),

\[
w_i = \sum_j a_{ij}v_j \geq 0.
\]

Also

\[
\sum_i w_i = \sum_i \sum_j a_{ij}v_j = \sum_j \sum_i a_{ij}v_j = \sum_j v_j = c.
\]
Theorem 12.1.6 Suppose $A$ is a Markov matrix in which $a_{ij} > 0$ for all $i, j$ and suppose $w$ is a vector. Then for each $i$,
\[
\lim_{k \to \infty} (A^k w)_i = v_i
\]
where $Av = v$. In words, $A^k w$ always converges to a steady state. In addition to this, if the vector $w$ satisfies $w_i \geq 0$ for all $i$ and $\sum_i w_i = c$, then the vector $v$ will also satisfy the conditions, $v_i \geq 0$, $\sum_i v_i = c$.

Proof: By Lemma 12.1.4, since each $a_{ij} > 0$, the eigenvalues are either 1 or have absolute value less than 1. Therefore, the claimed limit exists by Theorem 12.1.6. The assertion that the components are nonnegative and sum to $c$ follows from Lemma 12.1.4. That $Av = v$ follows from
\[
v = \lim_{n \to \infty} A^n w = \lim_{n \to \infty} A^{n+1} w = A \lim_{n \to \infty} A^n w = Av. \]

It is not hard to generalize the conclusion of this theorem to regular Markov processes.

Corollary 12.1.7 Suppose $A$ is a regular Markov matrix, one for which the entries of $A^k$ are all positive for some $k$, and suppose $w$ is a vector. Then for each $i$,
\[
\lim_{n \to \infty} (A^n w)_i = v_i
\]
where $Av = v$. In words, $A^n w$ always converges to a steady state. In addition to this, if the vector $w$ satisfies $w_i \geq 0$ for all $i$ and $\sum_i w_i = c$, then the vector $v$ will also satisfy the conditions $v_i \geq 0$, $\sum_i v_i = c$.

Proof: Let the entries of $A^k$ be all positive for some $k$. Now suppose that $a_{ij} \geq 0$ for all $i, j$ and $A = (a_{ij})$ is a Markov matrix. Then if $B = (b_{ij})$ is a Markov matrix with $b_{ij} > 0$ for all $ij$, it follows that $BA$ is a Markov matrix which has strictly positive entries. This is because the $ij^{th}$ entry of $BA$ is
\[
\sum_k b_{ik}a_{kj} > 0,
\]
Thus, from Lemma 12.1.4, $A^k$ has an eigenvalue equal to 1 for all $k$ sufficiently large, and all the other eigenvalues have absolute value strictly less than 1. The same must be true of $A$. If $v \neq 0$ and $Av = \lambda v$ and $|\lambda| = 1$, then $A^k v = \lambda^k v$ and so, by Lemma 12.1.4, $\lambda^m = 1$ if $m \geq k$. Thus
\[
1 = \lambda^{k+1} = \lambda^k \lambda = \lambda
\]
By Theorem 12.1.6, $\lim_{n \to \infty} A^n w$ exists. The rest follows as in Theorem 12.1.6.

12.2 Migration Matrices

Definition 12.2.1 Let $n$ locations be denoted by the numbers $1, 2, \ldots, n$. Also suppose it is the case that each year $a_{ij}$ denotes the proportion of residents in location $j$ which move to location $i$. Also suppose no one escapes or emigrates from without these $n$ locations. This last assumption requires $\sum_i a_{ij} = 1$. Thus $(a_{ij})$ is a Markov matrix referred to as a migration matrix.

If $v = (x_1, \ldots, x_n)^T$ where $x_i$ is the population of location $i$ at a given instant, you obtain the population of location $i$ one year later by computing $\sum_j a_{ij}x_j = (Av)_i$. Therefore, the population of location $i$ after $k$ years is $(A^k v)_i$. Furthermore, Corollary 12.1.7 can be used to predict in the case where $A$ is regular what the long time population will be for the given locations.

As an example of the above, consider the case where $n = 3$ and the migration matrix is of the form
\[
\begin{pmatrix}
.6 & 0 & .1 \\
.2 & .8 & 0 \\
.2 & .2 & .9
\end{pmatrix}.
\]
Now
\[
\begin{pmatrix} .6 & 0 & .1 \\ .2 & .8 & 0 \\ .2 & .2 & .9 \end{pmatrix}^2 = \begin{pmatrix} .38 & .02 & .15 \\ .28 & .64 & .02 \\ .34 & .34 & .83 \end{pmatrix}
\]
and so the Markov matrix is regular. Therefore, \((A^k \mathbf{v})_i\) will converge to the \(i^{th}\) component of a steady state. It follows the steady state can be obtained from solving the system
\[
\begin{align*}
.6x + .1z &= x \\
.2x + .8y &= y \\
.2x + .2y + .9z &= z
\end{align*}
\]
along with the stipulation that the sum of \(x, y,\) and \(z\) must equal the constant value present at the beginning of the process. The solution to this system is\(\{y = x, z = 4x, x = x\}\).

If the total population at the beginning is 150,000, then you solve the following system
\[
y = x, \ z = 4x, \ x + y + z = 150000
\]
whose solution is easily seen to be \(\{x = 25000, z = 100000, y = 25000\}\). Thus, after a long time there would be about four times as many people in the third location as in either of the other two.

### 12.3 Absorbing States

There is a different kind of Markov process containing so called absorbing states which result in transition matrices which are not regular. However, Theorem 12.1.3 may still apply. One such example is the Gambler’s ruin problem. There is a total amount of money denoted by \(b\). The Gambler starts with an amount \(j > 0\) and gambles till he either loses everything or gains everything. He does this by playing a game in which he wins with probability \(p\) and loses with probability \(q\). When he wins, the amount of money he has increases by 1 and when he loses, the amount of money he has decreases by 1. Thus the states are the integers from 0 to \(b\). Let \(p_{ij}\) denote the probability that the gambler has \(i\) at the end of a game given that he had \(j\) at the beginning. Let \(p_{ij}^n\) denote the probability that the gambler has \(i\) after \(n\) games given that he had \(j\) initially. Thus
\[
p_{ij}^{n+1} = \sum_k p_{ik}p_{kj}^n,
\]
and so \(p_{ij}^n\) is the \(ij^{th}\) entry of \(P^n\) where \(P\) is the transition matrix. The above description indicates that this transition probability matrix is of the form
\[
P = \begin{pmatrix} 1 & q & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ 0 & p & \ddots & q & \vdots \\ \vdots & \ddots & 0 & 0 & \vdots \\ 0 & \cdots & 0 & p & 1 \end{pmatrix}
\tag{12.2}
\]

The absorbing states are 0 and \(b\). In the first, the gambler has lost everything and hence has nothing else to gamble, so the process stops. In the second, he has won everything and there is nothing else to gain, so again the process stops.

Consider the eigenvalues of this matrix.
Lemma 12.3.1 Let \( p, q > 0 \) and \( p + q = 1 \). Then the eigenvalues of
\[
\begin{pmatrix}
0 & q & 0 & \cdots & 0 \\
p & 0 & q & \cdots & 0 \\
0 & p & 0 & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & q \\
0 & \cdots & \cdots & 0 & p
\end{pmatrix}
\]
have absolute value less than 1.

Proof: By Gerschgorin’s theorem, (See Page 334) if \( \lambda \) is an eigenvalue, then \( |\lambda| \leq 1 \). Now suppose \( \mathbf{v} \) is an eigenvector for \( \lambda \). Then
\[
A \mathbf{v} = \begin{pmatrix}
qv_2 \\
pv_1 + qv_3 \\
\vdots \\
pv_n - 2 + qv_n \\
pv_n - 1
\end{pmatrix} = \lambda \begin{pmatrix}
v_1 \\
v_2 \\
\vdots \\
v_{n-1} \\
v_n
\end{pmatrix}.
\]
Suppose \( |\lambda| = 1 \). Let \( v_k \) be the first nonzero entry. Then
\[
qv_{k+1} = \lambda v_k
\]
and so \( |v_{k+1}| > |v_k| \). If \( \{|v_j|\}_{j=k}^m \) is increasing for some \( m > k \), then
\[
p|v_{m-1}| + q|v_m| \geq |pv_{m-2} + qv_m| = |\lambda v_{m-1}| = |v_{m-1}|
\]
and so \( q|v_m| \geq q|v_{m-1}| \). Thus by induction, the sequence is increasing. Hence \( |v_n| \geq |v_{n-1}| > 0 \).
However, the last line states that \( p|v_{n-1}| = |v_n| \) which requires that \( |v_{n-1}| > |v_n| \), a contradiction.

Now consider the eigenvalues of \( \begin{pmatrix} 1 & \lambda & q & 0 & \cdots & 0 \\ 0 & -\lambda & \ddots & 0 \\ \vdots & \ddots & \ddots & q \\ 0 & \cdots & -\lambda & 0 \\ 0 & \cdots & 0 & p & 1 - \lambda \end{pmatrix} \). For \( P \) given there,
\[
P - \lambda I = \begin{pmatrix}
1 - \lambda & q & 0 & \cdots & 0 \\
0 & -\lambda & \ddots & 0 \\
0 & p & \ddots & q & \vdots \\
\vdots & \ddots & \ddots & -\lambda & 0 \\
0 & \cdots & 0 & p & 1 - \lambda
\end{pmatrix}
\]
and so, expanding the determinant of the matrix along the first column and then along the last column yields
\[
(1 - \lambda)^2 \det \begin{pmatrix}
-\lambda & q \\
p & \ddots & \ddots \\
\vdots & \ddots & -\lambda & q \\
p & \cdots & & -\lambda \\
p & \cdots & \cdots & p
\end{pmatrix}.
\]
The roots of the polynomial after \( (1 - \lambda)^2 \) have absolute value less than 1 because they are just the eigenvalues of a matrix of the sort in Lemma 12.3.1. It follows that the conditions of Theorem 12.1.3 apply and therefore, \( \lim_{n \to \infty} P^n \) exists.

Of course, the above transition matrix, models many other kinds of problems. It is called a Markov process with two absorbing states, sometimes a random walk with two absorbing states.
It is interesting to find the probability that the gambler loses all his money. This is given by 
\[ \lim_{n \to \infty} p^n_{0j} \]. From the transition matrix for the gambler’s ruin problem, it follows that 
\[ p^n_{0j} = \sum_k p^n_{0k} p_{kj} = q^n_{0(j-1)} + p^n_{0(j+1)} \]
for \( j \in [1, b - 1] \), 
\[ p^n_{00} = 1, \text{ and } p^n_{0b} = 0. \]
Assume here that \( p \neq q \). Now it was shown above that \( \lim_{n \to \infty} p^n_{0j} \) exists. Denote by \( P_j \) this limit. Then the above becomes much simpler if written as 
\[ P_j = qP_{j-1} + pP_{j+1} \]
for \( j \in [1, b - 1] \); 
\[ P_0 = 1 \text{ and } P_b = 0. \]
(12.3)
(12.4)

It is only required to find a solution to the above difference equation with boundary conditions. To do this, look for a solution in the form \( P_j = \frac{q^j}{p^j} \) and use the difference equation with boundary conditions to find the correct values of \( r \). Thus you need 
\[ r^j = qr^{j-1} + pr^{j+1} \]
and so to find \( r \) you need to have \( pr^2 - r + q = 0 \), and so the solutions for \( r \) are \( r = \frac{1}{2p} \left( 1 + \sqrt{1 - 4pq} \right), \frac{1}{2p} \left( 1 - \sqrt{1 - 4pq} \right) \)
Now 
\[ \sqrt{1 - 4pq} = \sqrt{1 - 4p(1 - p)} = \sqrt{1 - 4p + 4p^2} = 1 - 2p. \]
Thus the two values of \( r \) simplify to 
\[ \frac{1}{2p} \left( 1 + 1 - 2p \right) = \frac{q}{p}, \frac{1}{2p} \left( 1 - (1 - 2p) \right) = 1 \]
Therefore, for any choice of \( C_i, i = 1, 2, \)
\[ C_1 + C_2 \left( \frac{q}{p} \right)^j \]
will solve the difference equation. Now choose \( C_1, C_2 \) to satisfy the boundary conditions \[ P_0 \text{ and } P_b \]. Thus you need to have 
\[ C_1 + C_2 = 1, \ C_1 + C_2 \left( \frac{q}{p} \right)^b = 0 \]
It follows that 
\[ C_2 = \frac{q^b}{p^b - q^b}, \ C_1 = \frac{q^b}{q^b - p^b} \]
Thus \( P_j = \)
\[ \frac{q^b}{q^b - p^b} + \frac{p^b}{p^b - q^b} \left( \frac{q}{p} \right)^j = \frac{q^b}{q^b - p^b} - \frac{p^b}{p^b - q^b} = \frac{q^j (q^{b-j} - p^{b-j})}{q^b - p^b} \]
To find the solution in the case of a fair game, one could take the \( \lim_{p \to 1/2} \) of the above solution. Taking this limit, you get 
\[ P_j = \frac{b - j}{b}. \]
You could also verify directly in the case where \( p = q = 1/2 \) in \[ \text{12.3} ] \] and \[ \text{12.4} ] \] that \( P_j = 1 \) and \( P_j = j \) are two solutions to the difference equation and proceeding as before.
12.4 Positive Matrices

Earlier theorems about Markov matrices were presented. These were matrices in which all the entries were nonnegative and either the columns or the rows added to 1. It turns out that many of the theorems presented can be generalized to positive matrices. When this is done, the resulting theory is mainly due to Perron and Frobenius. I will give an introduction to this theory here following Karlin and Taylor [21].

**Definition 12.4.1** For $A$ a matrix or vector, the notation, $A >> 0$ will mean every entry of $A$ is positive. By $A > 0$ is meant that every entry is nonnegative and at least one is positive. By $A \geq 0$ is meant that every entry is nonnegative. Thus the matrix or vector consisting only of zeros is $\geq 0$. An expression like $A >> B$ will mean $A - B > 0$ with similar modifications for $>$ and $\geq$.

For the sake of this section only, define the following for $x = (x_1,\cdots,x_n)^T$, a vector.

$$|x| \equiv (|x_1|,\cdots,|x_n|)^T.$$  

Thus $|x|$ is the vector which results by replacing each entry of $x$ with its absolute value. Also define for $x \in \mathbb{C}^n$,

$$||x||_1 \equiv \sum_k |x_k|.$$  

**Lemma 12.4.2** Let $A >> 0$ and let $x > 0$. Then $Ax >> 0$.

**Proof:** $(Ax)_i = \sum_j A_{ij}x_j > 0$ because all the $A_{ij} > 0$ and at least one $x_j > 0$.

**Lemma 12.4.3** Let $A >> 0$. Define

$$S \equiv \{ \lambda : Ax > \lambda x \text{ for some } x >> 0 \},$$  

and let

$$K \equiv \{ x \geq 0 \text{ such that } ||x||_1 = 1 \}.$$  

Now define

$$S_1 \equiv \{ \lambda : Ax \geq \lambda x \text{ for some } x \in K \}.$$  

Then

$$\sup (S) = \sup (S_1).$$

**Proof:** Let $\lambda \in S$. Then there exists $x >> 0$ such that $Ax > \lambda x$. Consider $y \equiv x/||x||_1$. Then $||y||_1 = 1$ and $Ay > \lambda y$. Therefore, $\lambda \in S_1$ and so $S \subseteq S_1$. Therefore, $\sup (S) \leq \sup (S_1)$.

Now let $\lambda \in S_1$. Then there exists $x \geq 0$ such that $||x||_1 = 1$ so $x > 0$ and $Ax > \lambda x$. Letting $y \equiv Ax$, it follows from Lemma 12.4.2 that $Ay >> \lambda y$ and $y >> 0$. Thus $\lambda \in S$ and so $S_1 \subseteq S$ which shows that $\sup (S_1) \leq \sup (S)$.

This lemma is significant because the set, $\{ x \geq 0 \text{ such that } ||x||_1 = 1 \} \equiv K$ is a compact set in $\mathbb{R}^n$. Define

$$\lambda_0 \equiv \sup (S) = \sup (S_1).$$

The following theorem is due to Perron.

**Theorem 12.4.4** Let $A >> 0$ be an $n \times n$ matrix and let $\lambda_0$ be given in (12.5). Then

1. $\lambda_0 > 0$ and there exists $x_0 >> 0$ such that $Ax_0 = \lambda_0 x_0$ so $\lambda_0$ is an eigenvalue for $A$.

2. If $Ax = \mu x$ where $x \neq 0$, and $\mu \neq \lambda_0$. Then $|\mu| < \lambda_0$.

3. The eigenspace for $\lambda_0$ has dimension 1.

---

\(^1\)This notation is just about the most abominable thing imaginable because it is the same notation but entirely different meaning than the norm. However, it saves space in the presentation of this theory of positive matrices and avoids the use of new symbols. Please forget about it when you leave this section.
Proof: To see \( \lambda_0 > 0 \), consider the vector, \( \mathbf{e} \equiv (1, \cdots, 1)^T \). Then

\[
(A\mathbf{e})_i = \sum_j A_{ij} > 0
\]

and so \( \lambda_0 \) is at least as large as

\[
\min_i \sum_j A_{ij}.
\]

Let \( \{\lambda_k\} \) be an increasing sequence of numbers from \( S_1 \) converging to \( \lambda_0 \). Letting \( \mathbf{x}_k \) be the vector from \( K \) which occurs in the definition of \( \mathbf{S}_1 \), these vectors are in a compact set. Therefore, there exists a subsequence, still denoted by \( \mathbf{x}_k \) such that \( \mathbf{x}_k \to \mathbf{x}_0 \in K \) and \( \lambda_k \to \lambda_0 \). Then passing to the limit,

\[
A\mathbf{x}_0 \geq \lambda_0 \mathbf{x}_0, \quad \mathbf{x}_0 > 0.
\]

If \( A\mathbf{x}_0 > \lambda_0 \mathbf{x}_0 \), then letting \( \mathbf{y} \equiv A\mathbf{x}_0 \), it follows from Lemma 12.4.3 that \( Ay \gg \lambda_0 y \) and \( y \gg 0 \). But this contradicts the definition of \( \lambda_0 \) as the supremum of the elements of \( S \) because since \( Ay \gg \lambda_0 y \), it follows \( Ay \gg (\lambda_0 + \varepsilon) y \) for \( \varepsilon \) a small positive number. Therefore, \( A\mathbf{x}_0 = \lambda_0 \mathbf{x}_0 \). It remains to verify that \( \mathbf{x}_0 \gg 0 \). But this follows immediately from

\[
0 < \sum_j A_{ij}x_{0j} = (A\mathbf{x}_0)_i = \lambda_0 x_{0i}.
\]

This proves \( \Box \).

Next suppose \( A\mathbf{x} = \mu \mathbf{x} \) and \( \mathbf{x} \neq 0 \) and \( \mu \neq \lambda_0 \). Then \( |A\mathbf{x}| = |\mu| |\mathbf{x}| \). But this implies \( A |\mathbf{x}| \geq |\mu| |\mathbf{x}| \). (See the above abominable definition of \( |\mathbf{x}| \).

Case 1: \( |\mathbf{x}| \neq \mathbf{x} \) and \( |\mathbf{x}| \neq -\mathbf{x} \).

In this case, \( A |\mathbf{x}| > |A\mathbf{x}| = |\mu| |\mathbf{x}| \) and letting \( \mathbf{y} = A |\mathbf{x}| \), it follows \( \mathbf{y} \gg 0 \) and \( Ay \gg |\mu| y \) which shows \( Ay \gg (|\mu| + \varepsilon) y \) for sufficiently small positive \( \varepsilon \) and verifies \( |\mu| < \lambda_0 \).

Case 2: \( |\mathbf{x}| = \mathbf{x} \) or \( |\mathbf{x}| = -\mathbf{x} \).

In this case, the entries of \( \mathbf{x} \) are all real and have the same sign. Therefore, \( A |\mathbf{x}| = |A\mathbf{x}| = |\mu| |\mathbf{x}| \). Now let \( \mathbf{y} \equiv |\mathbf{x}| / |\mathbf{x}| \). Then \( Ay = |\mu| y \) and so \( |\mu| \in S_1 \) showing that \( |\mu| \leq \lambda_0 \). But also, the fact the entries of \( \mathbf{x} \) have the same sign shows \( \mu = |\mu| \) and so \( \mu \in S_1 \). Since \( \mu \neq \lambda_0 \), it must be that \( \mu = |\mu| < \lambda_0 \). This proves \( \Box \).

It remains to verify \( \Box \). Suppose then that \( Ay = \lambda_0 y \) and for all scalars \( \alpha, \alpha \mathbf{x}_0 \neq \mathbf{y} \). Then

\[
A \Re \mathbf{y} = \lambda_0 \Re \mathbf{y}, \quad A \Im \mathbf{y} = \lambda_0 \Im \mathbf{y}.
\]

If \( \Re \mathbf{y} = \alpha_1 \mathbf{x}_0 \) and \( \Im \mathbf{y} = \alpha_2 \mathbf{x}_0 \) for real numbers, \( \alpha_i \), then \( \mathbf{y} = (\alpha_1 + i\alpha_2) \mathbf{x}_0 \) and it is assumed this does not happen. Therefore, either

\[
t \Re \mathbf{y} \neq \mathbf{x}_0 \quad \text{for all } t \in \mathbb{R}
\]

or

\[
t \Im \mathbf{y} \neq \mathbf{x}_0 \quad \text{for all } t \in \mathbb{R}.
\]

Assume the first holds. Then varying \( t \in \mathbb{R} \), there exists a value of \( t \) such that \( \mathbf{x}_0 + t \Re \mathbf{y} \gg 0 \) but it is not the case that \( \mathbf{x}_0 + t \Re \mathbf{y} \gg 0 \). Then \( A (\mathbf{x}_0 + t \Re \mathbf{y}) \gg 0 \) by Lemma 12.4.3. But this implies \( \lambda_0 (\mathbf{x}_0 + t \Re \mathbf{y}) \gg 0 \) which is a contradiction. Hence there exist real numbers, \( \alpha_1 \) and \( \alpha_2 \) such that \( \Re \mathbf{y} = \alpha_1 \mathbf{x}_0 \) and \( \Im \mathbf{y} = \alpha_2 \mathbf{x}_0 \) showing that \( \mathbf{y} = (\alpha_1 + i\alpha_2) \mathbf{x}_0 \). This proves \( \Box \).

It is possible to obtain a simple corollary to the above theorem.

**Corollary 12.4.5** If \( A > 0 \) and \( A^m \gg 0 \) for some \( m \in \mathbb{N} \), then all the conclusions of the above theorem hold.

Proof: There exists \( \mu_0 > 0 \) such that \( A^m \mathbf{y}_0 = \mu_0 \mathbf{y}_0 \) for \( \mathbf{y}_0 \gg 0 \) by Theorem 12.4.3 and

\[
\mu_0 = \sup \{ \mu : A^m \mathbf{x} \geq \mu \mathbf{x} \text{ for some } \mathbf{x} \in K \}.
\]
Let $\lambda_0^m = \mu_0$. Then
\[
(A - \lambda_0 I) \left( A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I \right) y_0 = (A^m - \lambda_0^m I) y_0 = 0
\]
and so letting $x_0 = \left( A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I \right) y_0$, it follows $x_0 \gg 0$ and $Ax_0 = \lambda_0 x_0$.

Suppose now that $Ax = \mu x$ for $x \neq 0$ and $\mu \neq \lambda_0$. Suppose $|\mu| \geq \lambda_0$. Multiplying both sides by $A$, it follows $A^m x = \mu^m x$ and $|\mu|^m = |\mu|^m \geq |\mu|^m = \mu_0$ and so from Theorem 12.4.2, since $|\mu|^m \geq \mu_0$, and $\mu^m$ is an eigenvalue of $A^m$, it follows that $\mu^m = \mu_0$. But by Theorem 12.4.4 again, this implies $x = cy_0$ for some scalar, $c$ and hence $Ay_0 = \mu y_0$. Since $y_0 \gg 0$, it follows $\mu \geq 0$ and so $\mu = \lambda_0$, a contradiction. Therefore, $|\mu| < \lambda_0$.

Finally, if $Ax = \lambda_0 x$, then $A^m x = \lambda_0^m x$ and so $x = cy_0$ for some scalar, $c$. Consequently,
\[
(A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I) x = c \left( A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I \right) y_0 = cx_0.
\]
Hence
\[
m\lambda_0^{m-1} x = cx_0
\]
which shows the dimension of the eigenspace for $\lambda_0$ is one.

The following corollary is an extremely interesting convergence result involving the powers of positive matrices.

**Corollary 12.4.6** Let $A > 0$ and $A^m \gg 0$ for some $m \in \mathbb{N}$. Then for $\lambda_0$ given in 12.4.3, there exists a rank one matrix $P$ such that
\[
\lim_{m \to \infty} \left\| \left( \frac{A}{\lambda_0} \right)^m - P \right\| = 0.
\]

**Proof:** Considering $A^T$, and the fact that $A$ and $A^T$ have the same eigenvalues, Corollary 12.4.3 implies the existence of a vector, $v > 0$ such that
\[
A^T v = \lambda_0 v.
\]
Also let $x_0$ denote the vector such that $Ax_0 = \lambda_0 x_0$ with $x_0 \gg 0$. First note that $x_0^T v > 0$ because both these vectors have all entries positive. Therefore, $v$ may be scaled such that
\[
v^T x_0 = x_0^T v = 1.
\]
Define
\[
P \equiv x_0 v^T.
\]
Thanks to 12.4.6,
\[
\frac{A}{\lambda_0} P = x_0 v^T = P, \quad P \left( \frac{A}{\lambda_0} \right) = x_0 v^T \left( \frac{A}{\lambda_0} \right) = x_0 v^T = P,
\]
and
\[
P^2 = x_0 v^T x_0 v^T = v^T x_0 = P.
\]
Therefore,
\[
\left( \frac{A}{\lambda_0} - P \right)^2 = \left( \frac{A}{\lambda_0} \right)^2 - 2 \left( \frac{A}{\lambda_0} \right) P + P^2 = \left( \frac{A}{\lambda_0} \right)^2 - P.
\]
Continuing this way, using 12.4.6 repeatedly, it follows
\[
\left( \left( \frac{A}{\lambda_0} - P \right)^m = \left( \frac{A}{\lambda_0} \right)^m - P. \quad (12.9)
\]

The eigenvalues of \( \left( \frac{A}{\lambda_0} \right)^m - P \) are of interest because it is powers of this matrix which determine the convergence of \( \left( \frac{A}{\lambda_0} \right)^m \) to \( P \). Therefore, let \( \mu \) be a nonzero eigenvalue of this matrix. Thus

\[
\left( \left( \frac{A}{\lambda_0} \right) - P \right)x = \mu x \tag{12.10}
\]

for \( x \neq 0 \), and \( \mu \neq 0 \). Applying \( P \) to both sides and using the second formula of \( \ref{eq12.9} \) yields

\[
0 = (P - P)x = \left( P \left( \frac{A}{\lambda_0} \right) - P^2 \right)x = \mu Px.
\]

But since \( Px = 0 \), it follows from \( \ref{eq12.9} \) that

\[
A x = \lambda_0 \mu x
\]

which implies \( \lambda_0 \mu \) is an eigenvalue of \( A \). Therefore, by Corollary \( \ref{cor12.4.5} \) it follows that either \( \lambda_0 \mu = \lambda_0 \) in which case \( \mu = 1 \), or \( \lambda_0 |\mu| < \lambda_0 \) which implies \( |\mu| < 1 \). But if \( \mu = 1 \), then \( x \) is a multiple of \( x_0 \) and \( \ref{eq12.9} \) would yield

\[
\left( \left( \frac{A}{\lambda_0} \right) - P \right)x_0 = x_0
\]

which says \( x_0 - x_0 v^T x_0 = x_0 \) and so by \( \ref{eq12.9} \), \( x_0 = 0 \) contrary to the property that \( x_0 > 0 \). Therefore, \( |\mu| < 1 \) and so this has shown that the absolute values of all eigenvalues of \( \left( \frac{A}{\lambda_0} \right) - P \) are less than 1. By Gelfand’s theorem, Theorem \( \ref{thm15.2.3} \), it follows

\[
\left\| \left( \left( \frac{A}{\lambda_0} \right) - P \right)^m \right\|^{1/m} < r < 1
\]

whenever \( m \) is large enough. Now by \( \ref{eq12.9} \) this yields

\[
\left\| \left( \frac{A}{\lambda_0} \right)^m - P \right\| = \left\| \left( \left( \frac{A}{\lambda_0} \right) - P \right)^m \right\| \leq r^m
\]

whenever \( m \) is large enough. It follows

\[
\lim_{m \to \infty} \left\| \left( \frac{A}{\lambda_0} \right)^m - P \right\| = 0
\]

as claimed.

What about the case when \( A > 0 \) but maybe it is not the case that \( A > 0 \)? As before,

\[
K \equiv \{ x \geq 0 \text{ such that } ||x||_1 = 1 \}.
\]

Now define

\[
S_1 \equiv \{ \lambda : Ax \geq \lambda x \text{ for some } x \in K \}
\]

and

\[
\lambda_0 \equiv \sup S_1 \tag{12.11}
\]

**Theorem 12.4.7** Let \( A > 0 \) and let \( \lambda_0 \) be defined in \( \ref{cor12.4.7} \). Then there exists \( x_0 > 0 \) such that \( Ax_0 = \lambda_0 x_0 \).

**Proof:** Let \( E \) consist of the matrix which has a one in every entry. Then from Theorem \( \ref{thm12.4.6} \) it follows there exists \( x_0 \) such that \( \sup \{ \lambda : (A + \delta E)x \geq \lambda x \text{ for some } x \in K \} \).
Now if \( \alpha < \delta \)
\[
\{ \lambda : (A + \alpha E) \mathbf{x} \geq \lambda \mathbf{x} \text{ for some } \mathbf{x} \in K \} \subseteq \\
\{ \lambda : (A + \delta E) \mathbf{x} \geq \lambda \mathbf{x} \text{ for some } \mathbf{x} \in K \}
\]
and so \( \lambda_{0\delta} \geq \lambda_{0\alpha} \) because \( \lambda_{0\delta} \) is the sup of the second set and \( \lambda_{0\alpha} \) is the sup of the first. It follows
the limit, \( \lambda_1 \equiv \lim_{\delta \to 0^+} \lambda_{0\delta} \) exists. Taking a subsequence and using the compactness of \( K \), there
exists a subsequence, still denoted by \( \xi \) such that as \( \delta \to 0 \), \( \mathbf{x}_\delta \to \mathbf{x} \in K \). Therefore,

\[
A \mathbf{x} = \lambda_1 \mathbf{x}
\]
and so, in particular, \( A \mathbf{x} \geq \lambda_1 \mathbf{x} \) and so \( \lambda_1 \leq \lambda_0 \). But also, if \( \lambda \leq \lambda_0 \),

\[
\lambda \mathbf{x} \leq A \mathbf{x} < (A + \delta E) \mathbf{x}
\]
showing that \( \lambda_{0\delta} \geq \lambda \) for all such \( \lambda \). But then \( \lambda_{0\delta} \geq \lambda_0 \) also. Hence \( \lambda_1 \geq \lambda_0 \), showing these two
numbers are the same. Hence \( A \mathbf{x} = \lambda_0 \mathbf{x} \).

If \( A^m \gg 0 \) for some \( m \) and \( A > 0 \), it follows that the dimension of the eigenspace for \( \lambda_0 \) is
one and that the absolute value of any other eigenvalue of \( A \) is less than \( \lambda_0 \). If it is only assumed
that \( A > 0 \), not necessarily \( \gg 0 \), this is no longer true. However, there is something which is very
interesting which can be said. First here is an interesting lemma.

**Lemma 12.4.8** Let \( M \) be a matrix of the form

\[
M = \begin{pmatrix} A & 0 \\ B & C \end{pmatrix}
\]
or

\[
M = \begin{pmatrix} A & B \\ 0 & C \end{pmatrix}
\]
where \( A \) is an \( r \times r \) matrix and \( C \) is an \( (n-r) \times (n-r) \) matrix. Then \( \det(M) = \det(A) \det(B) \)
and \( \sigma(M) = \sigma(A) \cup \sigma(C) \).

**Proof:** To verify the claim about the determinants, note

\[
\begin{pmatrix} A & 0 \\ B & C \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} \left( \begin{pmatrix} I & 0 \\ B & C \end{pmatrix} \right)
\]
Therefore,

\[
\det \begin{pmatrix} A & 0 \\ B & C \end{pmatrix} = \det \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} \det \begin{pmatrix} I & 0 \\ B & C \end{pmatrix}.
\]
But it is clear from the method of Laplace expansion that

\[
\det \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} = \det A
\]
and from the multilinear properties of the determinant and row operations that

\[
\det \begin{pmatrix} I & 0 \\ B & C \end{pmatrix} = \det \begin{pmatrix} I & 0 \\ 0 & C \end{pmatrix} = \det C.
\]
The case where \( M \) is upper block triangular is similar.

This immediately implies \( \sigma(M) = \sigma(A) \cup \sigma(C) \).

**Theorem 12.4.9** Let \( A > 0 \) and let \( \lambda_0 \) be given in (12.11). If \( \lambda \) is an eigenvalue for \( A \) such that
\( |\lambda| = \lambda_0 \), then \( \lambda/\lambda_0 \) is a root of unity. Thus \( (\lambda/\lambda_0)^m = 1 \) for some \( m \in \mathbb{N} \).
Proof: Applying Theorem 12.4.7 to $A^T$, there exists $v > 0$ such that $A^T v = \lambda_0 v$. In the first part of the argument it is assumed $v \gg 0$. Now suppose $A x = \lambda x, x \neq 0$ and that $|\lambda| = \lambda_0$. Then

$$A |x| \geq |\lambda| |x| = \lambda_0 |x|$$

and it follows that if $A |x| > |\lambda| |x|$, then since $v \gg 0$,

$$\lambda_0 \langle v, |x| \rangle < \langle v, A |x| \rangle = (A^T v, |x|) = \lambda_0 \langle v, |x| \rangle,$$

a contradiction. Therefore,

$$A |x| = \lambda_0 |x|.$$  \hfill (12.12)

It follows that

$$\sum_j A_{ij} x_j \lambda_0 |x_i| = \sum_j A_{ij} |x_j|$$

and so the complex numbers,

$$A_{ij} x_j, A_{ik} x_k$$

must have the same argument for every $k,j$ because equality holds in the triangle inequality. Therefore, there exists a complex number, $\mu_i$ such that

$$A_{ij} x_j = \mu_i A_{ij} |x_j|$$  \hfill (12.13)

and so, letting $r \in \mathbb{N}$,

$$A_{ij} x_j \mu_j^r = \mu_i A_{ij} |x_j| \mu_j^r.$$

Summing on $j$ yields

$$\sum_j A_{ij} x_j \mu_j^r = \mu_i \sum_j A_{ij} |x_j| \mu_j^r.$$  \hfill (12.14)

Also, summing on $j$ and using that $\lambda$ is an eigenvalue for $x$, it follows from (12.13) that

$$\lambda x_i = \sum_j A_{ij} x_j = \mu_i \sum_j A_{ij} |x_j| = \mu_i \lambda_0 |x_i|.$$  \hfill (12.15)

From (12.14) and (12.15),

$$\sum_j A_{ij} x_j \mu_j^r = \mu_i \sum_j A_{ij} |x_j| \mu_j^r$$

$$= \mu_i \sum_j A_{ij} \mu_j |x_j| \mu_j^{r-1}$$

$$= \mu_i \sum_j A_{ij} \left( \frac{\lambda}{\lambda_0} \right) x_j \mu_j^{r-1}$$

$$= \mu_i \left( \frac{\lambda}{\lambda_0} \right) \sum_j A_{ij} x_j \mu_j^{r-1}$$

Now from (12.15) with $r$ replaced by $r - 1$, this equals

$$\mu_i^2 \left( \frac{\lambda}{\lambda_0} \right) \sum_j A_{ij} |x_j| \mu_j^{r-1} = \mu_i^2 \left( \frac{\lambda}{\lambda_0} \right) \sum_j A_{ij} |x_j| \mu_j^{r-2}$$

$$= \mu_i^2 \left( \frac{\lambda}{\lambda_0} \right)^2 \sum_j A_{ij} x_j \mu_j^{r-2}.$$
Continuing this way,
\[ \sum_j A_{ij} x_j \mu_j = \mu_i \left( \frac{\lambda}{\lambda_0} \right)^k \sum_j A_{ij} x_j \mu_j^{r-k} \]
and eventually, this shows
\[ \sum_j A_{ij} x_j \mu_j = \mu_i \left( \frac{\lambda}{\lambda_0} \right)^r \sum_j A_{ij} x_j \]
\[ = \left( \frac{\lambda}{\lambda_0} \right)^r \lambda (x_i \mu_i^r) \]
and this says \( \left( \frac{\lambda}{\lambda_0} \right)^{r+1} \) is an eigenvalue for \( \left( \frac{A}{\lambda_0} \right) \) with the eigenvector being
\[ (x_1 \mu_1^r, \ldots, x_n \mu_n^r)^T. \]

Now recall that \( r \in \mathbb{N} \) was arbitrary and so this has shown that \( \left( \frac{\lambda}{\lambda_0} \right)^2, \left( \frac{\lambda}{\lambda_0} \right)^3, \left( \frac{\lambda}{\lambda_0} \right)^4, \ldots \) are each eigenvalues of \( \left( \frac{A}{\lambda_0} \right) \) which has only finitely many and hence this sequence must repeat. Therefore, \( \left( \frac{\lambda}{\lambda_0} \right) \) is a root of unity as claimed. This proves the theorem in the case that \( v >> 0 \).

Now it is necessary to consider the case where \( v > 0 \) but it is not the case that \( v >> 0 \). Then in this case, there exists a permutation matrix \( P \) such that
\[
Pv = \begin{pmatrix} v_1 \\ \vdots \\ v_r \\ 0 \\ \vdots \\ 0 \end{pmatrix} \equiv \begin{pmatrix} u \\ 0 \end{pmatrix} = v_1.\]

Then
\[ \lambda_0 v = A^T v = A^T P v_1. \]

Therefore, \[ \lambda_0 v_1 = P A^T P v_1 = G v_1 \]

Now \( P^2 = I \) because it is a permutation matrix. Therefore, the matrix \( G \equiv PA^T P \) and \( A \) are similar. Consequently, they have the same eigenvalues and it suffices from now on to consider the matrix \( G \) rather than \( A \). Then
\[ \lambda_0 \begin{pmatrix} u \\ 0 \end{pmatrix} = \begin{pmatrix} M_1 & M_2 \\ M_3 & M_4 \end{pmatrix} \begin{pmatrix} u \\ 0 \end{pmatrix} \]

where \( M_1 = r \times r \) and \( M_4 = (n-r) \times (n-r) \). It follows from block multiplication and the assumption that \( A \) and hence \( G \) are \( > 0 \) that
\[ G = \begin{pmatrix} A' & B \\ 0 & C \end{pmatrix}. \]

Now let \( \lambda \) be an eigenvalue of \( G \) such that \( |\lambda| = \lambda_0 \). Then from Lemma [THEOREM], either \( \lambda \in \sigma(A') \) or \( \lambda \in \sigma(C) \). Suppose without loss of generality that \( \lambda \in \sigma(A') \). Since \( A' > 0 \) it has a largest positive eigenvalue \( \lambda'_0 \) which is obtained from [THEOREM]. Thus \( \lambda'_0 \leq \lambda_0 \) but \( \lambda \) being an eigenvalue of \( A' \), has its absolute value bounded by \( \lambda'_0 \) and so \( \lambda_0 = |\lambda| \leq \lambda'_0 \leq \lambda_0 \) showing that \( \lambda_0 \in \sigma(A') \). Now if there exists \( v >> 0 \) such that \( A'^T v = \lambda_0 v \), then the first part of this proof applies to the matrix \( A \).
and so \((\lambda/\lambda_0)\) is a root of unity. If such a vector, \(v\) does not exist, then let \(A'\) play the role of \(A\) in the above argument and reduce to the consideration of

\[
G' = \begin{pmatrix} \lambda & B' \\ 0 & C' \end{pmatrix}
\]

where \(G'\) is similar to \(A'\) and \(\lambda, \lambda_0 \in \sigma (A')\). Stop if \(A'^T v = \lambda_0 v\) for some \(v > 0\). Otherwise, decompose \(A''\) similar to the above and add another prime. Continuing this way you must eventually obtain the situation where \((A''')^T v = \lambda_0 v\) for some \(v > 0\). Indeed, this happens no later than when \(A''''\) is a 1 \(1\) matrix.

### 12.5 Functions Of Matrices

The existence of the Jordan form also makes it possible to define various functions of matrices. Suppose

\[
f(\lambda) = \sum_{n=0}^{\infty} a_n \lambda^n	ag{12.16}
\]

for all \(|\lambda| < R\). There is a formula for \(f(A) = \sum_{n=0}^{\infty} a_n A^n\) which makes sense whenever \(\rho (A) < R\). Thus you can speak of \(f(A)\) for \(A\) an \(n \times n\) matrix. To begin with, define

\[
f_P (\lambda) = \sum_{n=0}^{P} a_n \lambda^n
\]

so for \(k < P\)

\[
f_P^{(k)} (\lambda) = \sum_{n=k}^{P} a_n n \cdots (n-k+1) \lambda^{n-k} = \sum_{n=k}^{P} a_n \binom{n}{k} k! \lambda^{n-k}.	ag{12.17}
\]

Thus

\[
f_P^{(k)} (\lambda) = \sum_{n=k}^{P} a_n \binom{n}{k} k! \lambda^{n-k}\tag{12.18}
\]

To begin with consider \(f(J_m(\lambda))\) where \(J_m(\lambda)\) is an \(m \times m\) Jordan block. Thus \(J_m(\lambda) = D + N\) where \(N^m = 0\) and \(N\) commutes with \(D\). Therefore, letting \(P > m\)

\[
\sum_{n=0}^{P} a_n J_m(\lambda)^n = \sum_{n=0}^{P} a_n \sum_{k=0}^{n} \binom{n}{k} D^{n-k} N^k
\]

\[
= \sum_{k=0}^{m-1} \sum_{n=k}^{P} a_n \binom{n}{k} D^{n-k} N^k
\]

\[
= \sum_{k=0}^{m-1} N^k \sum_{n=k}^{P} a_n \binom{n}{k} D^{n-k}.	ag{12.19}
\]

From (12.18) this equals

\[
\sum_{k=0}^{m-1} N^k \text{diag} \left( \frac{f_P^{(k)} (\lambda)}{k!}, \cdots, \frac{f_P^{(k)} (\lambda)}{k!} \right)	ag{12.20}
\]

where for \(k = 0, \cdots, m-1\), define \(\text{diag}_k (a_1, \cdots, a_{m-k})\) the \(m \times m\) matrix which equals zero everywhere except on the \(k^{th}\) super diagonal where this diagonal is filled with the numbers, \(\{a_1, \cdots, a_{m-k}\}\)
from the upper left to the lower right. With no subscript, it is just the diagonal matrices having the
indicated entries. Thus in \(4 \times 4\) matrices, \(\text{diag}_2 (1, 2)\) would be the matrix
\[
\begin{pmatrix}
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 2 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}.
\]
Then from Eqs. 12.20 and 12.17,
\[
\begin{align*}
p_{\sum_{n=0}^{P} a_n J_m (\lambda)^n} &= \sum_{k=0}^{m-1} \text{diag} \left( \frac{f'(\lambda)}{k!}, \ldots, \frac{f^{(m-1)}(\lambda)}{(m-1)!} \right). \\
\end{align*}
\]
Therefore, \(\sum_{n=0}^{P} a_n J_m (\lambda)^n = \)
\[
\begin{pmatrix}
f_P(\lambda) & \frac{f'_P(\lambda)}{1!} & \frac{f''(\lambda)}{2!} & \cdots & \frac{f^{(m-1)}(\lambda)}{(m-1)!} \\
\frac{f'_P(\lambda)}{1!} & f_P(\lambda) & \frac{f''(\lambda)}{2!} & \cdots & \frac{f^{(m-1)}(\lambda)}{(m-1)!} \\
\vdots & \vdots & \ddots & \ddots & \ddots \\
0 & \frac{f''(\lambda)}{2!} & \cdots & f_P(\lambda) & \frac{f'_P(\lambda)}{1!} \\
0 & \frac{f'_P(\lambda)}{1!} & \cdots & \frac{f''(\lambda)}{2!} & f_P(\lambda)
\end{pmatrix}
\tag{12.21}
\]
Now let \(A\) be an \(n \times n\) matrix with \(\rho (A) < R\) where \(R\) is given above. Then the Jordan form of \(A\) is of the form
\[
J = \begin{pmatrix}
J_1 & & & \\
& J_2 & & \\
& & \ddots & \\
& & & J_r
\end{pmatrix}
\tag{12.22}
\]
where \(J_k = J_{m_k} (\lambda_k)\) is an \(m_k \times m_k\) Jordan block and \(A = S^{-1} JS\). Then, letting \(P > m_k\) for all \(k\),
\[
\sum_{n=0}^{P} a_n A^n = S^{-1} \sum_{n=0}^{P} a_n J^n S,
\]
and because of block multiplication of matrices,
\[
\sum_{n=0}^{P} a_n J^n = \begin{pmatrix}
\sum_{n=0}^{P} a_n J_1^n & & & 0 \\
& \ddots & & \\
& & \ddots & \\
0 & & & \sum_{n=0}^{P} a_n J_r^n
\end{pmatrix}
\]
and from Eqs. 12.21 \(\sum_{n=0}^{P} a_n J_k^n\) converges as \(P \to \infty\) to the \(m_k \times m_k\) matrix
\[
\begin{pmatrix}
f (\lambda_k) & \frac{f' (\lambda_k)}{1!} & \frac{f''(\lambda_k)}{2!} & \cdots & \frac{f^{(m-1)}(\lambda_k)}{(m-1)!} \\
0 & f (\lambda_k) & \frac{f'(\lambda_k)}{1!} & \cdots & \frac{f^{(m-1)}(\lambda_k)}{(m-1)!} \\
0 & 0 & f (\lambda_k) & \frac{f'(\lambda_k)}{1!} & \frac{f''(\lambda_k)}{2!} \\
\vdots & \vdots & \ddots & \ddots & \ddots \\
0 & 0 & \cdots & 0 & f (\lambda_k)
\end{pmatrix}
\tag{12.23}
\]
There is no convergence problem because \(|\lambda| < R\) for all \(\lambda \in \sigma (A)\). This has proved the following theorem.
Theorem 12.5.1 Let \( f \) be given by \((12.16)\) and suppose \( \rho(A) < R \) where \( R \) is the radius of convergence of the power series in \((12.16)\). Then the series,

\[
\sum_{k=0}^{\infty} a_n A^n
\]

(12.24)

converges in the space \( L(F^n, F^n) \) with respect to any of the norms on this space and furthermore,

\[
\sum_{k=0}^{\infty} a_n A^n = S^{-1}\begin{pmatrix}
\sum_{n=0}^{\infty} a_n J^n_1 & 0 \\
\vdots & \ddots & \ddots \\
0 & \cdots & \sum_{n=0}^{\infty} a_n J^n_r
\end{pmatrix} S
\]

where \( \sum_{n=0}^{\infty} a_n J^n_k \) is an \( m_k \times m_k \) matrix of the form given in \((12.23)\) where \( A = S^{-1}JS \) and the Jordan form of \( A \), \( J \) is given by \((12.22)\). Therefore, you can define \( f(A) \) by the series in \((12.24)\).

Here is a simple example.

Example 12.5.2 Find \( \sin(A) \) where \( A = \begin{pmatrix} 4 & 1 & -1 & 1 \\ 1 & 1 & 0 & -1 \\ 0 & -1 & 1 & -1 \\ -1 & 2 & 1 & 4 \end{pmatrix} \).

In this case, the Jordan canonical form of the matrix is not too hard to find.

\[
\begin{pmatrix} 4 & 1 & -1 & 1 \\ 1 & 1 & 0 & -1 \\ 0 & -1 & 1 & -1 \\ -1 & 2 & 1 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 0 & -2 & -1 \\ 1 & -4 & -2 & -1 \\ 0 & 0 & -2 & 1 \\ -1 & 4 & 4 & 2 \end{pmatrix}
\]

\[
\begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 & 0 & 1/2 \\ 1/8 & -3/8 & 0 & -1/8 \\ 0 & 1/4 & -3/4 & 1/4 \\ 0 & 1/2 & 1/2 & 1/2 \end{pmatrix}
\]

Then from the above theorem \( \sin(J) \) is given by

\[
\sin\begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} \sin 4 & 0 & 0 & 0 \\ 0 & \sin 2 & \cos 2 & -\sin 2/2 \\ 0 & 0 & \sin 2 & \cos 2 \\ 0 & 0 & 0 & \sin 2 \end{pmatrix}
\]

Therefore, \( \sin(A) =
\]

\[
\begin{pmatrix} 2 & 0 & -2 & -1 \\ 1 & -4 & -2 & -1 \\ 0 & 0 & -2 & 1 \\ -1 & 4 & 4 & 2 \end{pmatrix} \begin{pmatrix} \sin 4 & 0 & 0 & 0 \\ 0 & \sin 2 & \cos 2 & -\sin 2/2 \\ 0 & 0 & \sin 2 & \cos 2 \\ 0 & 0 & 0 & \sin 2 \end{pmatrix} \begin{pmatrix} 1/2 & 1/2 & 0 & 1/2 \\ 1/8 & -3/8 & 0 & -1/8 \\ 0 & 1/4 & -3/4 & 1/4 \\ 0 & 1/2 & 1/2 & 1/2 \end{pmatrix} = M
\]
where the columns of $M$ are as follows from left to right,

\[
\begin{pmatrix}
\sin 4 \\
\frac{1}{2}\sin 4 - \frac{1}{2}\sin 2 \\
0 \\
-\frac{1}{2}\sin 4 + \frac{1}{2}\sin 2
\end{pmatrix}
\begin{pmatrix}
\sin 4 - \sin 2 - \cos 2 \\
\frac{1}{2}\sin 4 + \frac{3}{2}\sin 2 - 2\cos 2 \\
-\cos 2 \\
-\frac{1}{2}\sin 4 - \frac{3}{2}\sin 2 + 3\cos 2
\end{pmatrix}
\begin{pmatrix}
-\cos 2 \\
\sin 2 \\
\sin 2 - \cos 2 \\
\cos 2 - \sin 2
\end{pmatrix}
\begin{pmatrix}
\sin 4 - \sin 2 - \cos 2 \\
\frac{1}{2}\sin 4 + \frac{1}{2}\sin 2 - 2\cos 2 \\
-\cos 2 \\
-\frac{1}{2}\sin 4 + \frac{1}{2}\sin 2 + 3\cos 2
\end{pmatrix}
\begin{pmatrix}
\sin 4 \\
\frac{1}{2}\sin 4 - \frac{1}{2}\sin 2 \\
0 \\
-\frac{1}{2}\sin 4 + \frac{1}{2}\sin 2
\end{pmatrix}.
\]

Perhaps this isn’t the first thing you would think of. Of course the ability to get this nice closed form description of $\sin (A)$ was dependent on being able to find the Jordan form along with a similarity transformation which will yield the Jordan form.

The following corollary is known as the spectral mapping theorem.

**Corollary 12.5.3** Let $A$ be an $n \times n$ matrix and let $\rho (A) < R$ where for $|\lambda| < R,$

\[f(\lambda) = \sum_{n=0}^{\infty} a_n \lambda^n.\]

Then $f(A)$ is also an $n \times n$ matrix and furthermore, $\sigma(f(A)) = f(\sigma(A))$. Thus the eigenvalues of $f(A)$ are exactly the numbers $f(\lambda)$ where $\lambda$ is an eigenvalue of $A$. Furthermore, the algebraic multiplicity of $f(\lambda)$ coincides with the algebraic multiplicity of $\lambda$.

All of these things can be generalized to linear transformations defined on infinite dimensional spaces and when this is done the main tool is the Dunford integral along with the methods of complex analysis. It is good to see it done for finite dimensional situations first because it gives an idea of what is possible.

### 12.6 Exercises

1. Suppose the migration matrix for three locations is

\[
\begin{pmatrix}
.5 & 0 & .3 \\
.3 & .8 & 0 \\
.2 & .2 & .7
\end{pmatrix}.
\]

Find a comparison for the populations in the three locations after a long time.

2. Show that if $\sum_i a_{ij} = 1$, then if $A = (a_{ij})$, then the sum of the entries of $Av$ equals the sum of the entries of $v$. Thus it does not matter whether $a_{ij} \geq 0$ for this to be so.

3. If $A$ satisfies the conditions of the above problem, can it be concluded that $\lim_{n \to \infty} A^n$ exists?

4. Give an example of a non regular Markov matrix which has an eigenvalue equal to $-1$.

5. Show that when a Markov matrix is non defective, all of the above theory can be proved very easily. In particular, prove the theorem about the existence of $\lim_{n \to \infty} A^n$ if the eigenvalues are either 1 or have absolute value less than 1.

6. Find a formula for $A^n$ where

\[
A = \begin{pmatrix}
\frac{5}{2} & -\frac{1}{2} & 0 & -1 \\
5 & 0 & 0 & -4 \\
\frac{7}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{5}{2} \\
\frac{7}{2} & -\frac{1}{2} & 0 & -2
\end{pmatrix}
\]
Does $\lim_{n \to \infty} A^n$ exist? Note that all the rows sum to 1. **Hint:** This matrix is similar to a diagonal matrix. The eigenvalues are $1, -1, \frac{1}{2}, \frac{1}{2}$.

7. Find a formula for $A^n$ where

\[
A = \begin{pmatrix}
2 & -\frac{1}{2} & \frac{1}{2} & -1 \\
4 & 0 & 1 & -4 \\
\frac{5}{2} & -\frac{1}{2} & 1 & -2 \\
3 & -\frac{1}{2} & \frac{1}{2} & -2
\end{pmatrix}
\]

Note that the rows sum to 1 in this matrix also. **Hint:** This matrix is not similar to a diagonal matrix but you can find the Jordan form and consider this in order to obtain a formula for this product. The eigenvalues are $1, \frac{1}{2}, \frac{1}{2}$.

8. Find $\lim_{n \to \infty} A^n$ if it exists for the matrix

\[
A = \begin{pmatrix}
\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 \\
-\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & 0 \\
\frac{1}{2} & \frac{1}{2} & \frac{3}{2} & 0 \\
\frac{1}{2} & \frac{3}{2} & \frac{3}{2} & 1
\end{pmatrix}
\]

The eigenvalues are $\frac{1}{2}, 1, 1, 1$.

9. Give an example of a matrix $A$ which has eigenvalues which are either equal to $1, -1$, or have absolute value strictly less than 1 but which has the property that $\lim_{n \to \infty} A^n$ does not exist.

10. If $A$ is an $n \times n$ matrix such that all the eigenvalues have absolute value less than 1, show $\lim_{n \to \infty} A^n = 0$.

11. Find an example of a $3 \times 3$ matrix $A$ such that $\lim_{n \to \infty} A^n$ does not exist but $\lim_{r \to \infty} A^{5r}$ does exist.

12. If $A$ is a Markov matrix and $B$ is similar to $A$, does it follow that $B$ is also a Markov matrix?

13. In Theorem [12.1.3] suppose everything is unchanged except that you assume either $\sum_i a_{ij} \leq 1$ or $\sum_j a_{ij} \leq 1$. Would the same conclusion be valid? What if you don’t insist that each $a_{ij} \geq 0$? Would the conclusion hold in this case?

14. Let $V$ be an $n$ dimensional vector space and let $x \in V$ and $x \neq 0$. Consider

\[\beta_x \equiv x, Ax, \ldots, A^{m-1}x\]

where

\[A^m x \in \text{span} (x, Ax, \ldots, A^{m-1}x)\]

and $m$ is the smallest such that the above inclusion in the span takes place. Show that $\{x, Ax, \ldots, A^{m-1}x\}$ must be linearly independent. Next suppose $\{v_1, \ldots, v_n\}$ is a basis for $V$. Consider $\beta_{v_i}$ as just discussed, having length $m_i$. Thus $A^{m_i} v_i$ is a linear combination of $v_i, Av_i, \ldots, A^{m-1} v_i$ for $m$ as small as possible. Let $p_{v_i}(\lambda)$ be the monic polynomial which expresses this linear combination. Thus $p_{v_i}(\lambda) v_i = 0$ and the degree of $p_{v_i}(\lambda)$ is as small as possible for this to take place. Show that the minimum polynomial for $A$ must be the monic polynomial which is the least common multiple of these polynomials $p_{v_i}(\lambda)$. 
Chapter 13

Inner Product Spaces

In this chapter is a more complete discussion of important theorems for inner product spaces. These results are presented for inner product spaces, the typical example being $\mathbb{C}^n$ or $\mathbb{R}^n$. The extra generality is used because most of the ideas have a straightforward generalization to something called a Hilbert space which is just a complete inner product space. First is a major result about projections.

13.1 Orthogonal Projections

If you have a closed subspace of an inner product space, it turns out that for a given point in the space, there is a closest point in this closed subspace.

Lemma 13.1.1 Suppose $\{u_j\}_{j=1}^n$ is an orthonormal basis for an inner product space $X$. Then for all $x \in X$,

$$x = \sum_{j=1}^n (x, u_j) u_j.$$  

Proof: Since $\{u_j\}_{j=1}^n$ is a basis, there exist unique scalars $\{\alpha_i\}$ such that

$$x = \sum_{j=1}^n \alpha_j u_j.$$  

It only remains to identify $\alpha_k$. From the properties of the inner product,

$$(x, u_k) = \sum_{j=1}^n \alpha_j (u_j, u_k) = \sum_{j=1}^n \alpha_j \delta_{jk} = \alpha_k \blacksquare$$

The following theorem is of fundamental importance. First note that a subspace of an inner product space is also an inner product space because you can use the same inner product.

Theorem 13.1.2 Let $M$ be a finite dimensional subspace of $X$, an inner product space and let $\{e_i\}_{i=1}^m$ be an orthonormal basis for $M$. Then if $y \in X$ and $w \in M$,

$$|y - w|^2 = \inf \left\{|y - z|^2 : z \in M \right\}$$ (13.1)

if and only if

$$(y - w, z) = 0$$ (13.2)

for all $z \in M$. Furthermore,

$$w = \sum_{i=1}^m (y, x_i) x_i$$ (13.3)

is the unique element of $M$ which has this property. It is called the orthogonal projection.
**Proof:** First we show that if (13.2), then (13.1). Let \( z \in M \) be arbitrary. Then

\[
|y - z|^2 = |y - w + (w - z)|^2
\]

\[
= (y - w + (w - z), y - w + (w - z))
\]

\[
= |y - w|^2 + |z - w|^2 + 2 \text{Re} (y - w, w - z)
\]

The last term is given to be 0 and so

\[
|y - z|^2 = |y - w|^2 + |z - w|^2
\]

which verifies (13.1).

Next suppose (13.1). Is it true that (13.2) follows? Let \( z \in M \) be arbitrary and let

\[
|x - w, w - z| = 1, \theta = (x - w, w - z) = |(x - w, w - z)|
\]

Then let

\[
p(t) = |x - w + t \theta (w - z)|^2 = |x - w|^2 + 2 \text{Re} (x - w, t \theta (w - z)) + t^2 |w - z|^2
\]

\[
= |x - w|^2 + 2 t \text{Re} \theta (x - w, (w - z)) + t^2 |w - z|^2
\]

\[
= |x - w|^2 + 2 t |(x - w, (w - z))| + t^2 |w - z|^2
\]

Then \( p \) has a minimum when \( t = 0 \) and so \( p'(0) = 2 \text{Re} (x - w, (w - z)) = 0 \) which shows (13.2). This proves the first part of the theorem since \( z \) is arbitrary.

It only remains to verify that \( w \) given in (13.3) satisfies (13.2) and is the only point of \( M \) which does so.

First, could there be two minimizers? Say \( w_1, w_2 \) both work. Then by the above characterization of minimizers,

\[
(x - w_1, w_1 - w_2) = 0
\]

\[
(x - w_2, w_1 - w_2) = 0
\]

Subtracting gives \( (w_1 - w_2, w_1 - w_2) = 0 \). Hence the minimizer is unique.

Finally, it remains to show that the given formula works. Letting \( \{e_1, \ldots, e_m\} \) be an orthonormal basis for \( M \), such a thing existing by the Gramm Schmidt process,

\[
(x - \sum_{i=1}^{m} (x, e_i) e_i, e_k) = (x, e_k) - \sum_{i=1}^{m} (x, e_i) (e_i, e_k)
\]

\[
= (x, e_k) - \sum_{i=1}^{m} (x, e_i) \delta_{ik}
\]

\[
= (x, e_k) - (x, e_k) = 0
\]

Since this inner product equals 0 for arbitrary \( e_k \), it follows that

\[
(x - \sum_{i=1}^{m} (x, e_i) e_i, z) = 0
\]

for every \( z \in M \) because each such \( z \) is a linear combination of the \( e_i \). Hence \( \sum_{i=1}^{m} (x, e_i) e_i \) is the unique minimizer.

**Example 13.1.3** Consider \( X \) equal to the continuous functions defined on \([-\pi, \pi]\) and let the inner product be given by

\[
\int_{-\pi}^{\pi} f(x) g(x) dx
\]
It is left to the reader to verify that this is an inner product. Letting $e_k$ be the function $x \to \frac{1}{\sqrt{2\pi}} e^{ikx}$, define

$$M \equiv \text{span} \left( \{e_k\}_{k=-n}^n \right).$$

Then you can verify that

$$\langle e_k, e_m \rangle = \int_{-\pi}^{\pi} \left( \frac{1}{\sqrt{2\pi}} e^{-ikx} \right) \left( \frac{1}{\sqrt{2\pi}} e^{imx} \right) dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i(m-k)x} = \delta_{km}$$

then for a given function $f \in X$, the function from $M$ which is closest to $f$ in this inner product norm is

$$g = \sum_{k=-n}^n \langle f, e_k \rangle e_k$$

In this case $\langle f, e_k \rangle = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} f(x) e^{ikx} dx$. These are the Fourier coefficients. The above is the $n$th partial sum of the Fourier series.

To show how this kind of thing approximates a given function, let $f(x) = x^2$. Let $M = \text{span} \left( \{\frac{1}{\sqrt{2\pi}} e^{-ikx}\}_{k=-3}^3 \right).$ Then, doing the computations, you find the closest point is of the form

$$\frac{1}{3} \sqrt{2\pi} \left( \frac{1}{\sqrt{2\pi}} \right) + \sum_{k=1}^3 \left( \frac{(-1)^k}{k^2} \right) \frac{1}{\sqrt{2\pi}} e^{-ikx} + \sum_{k=1}^3 \left( \frac{(-1)^k}{k^2} \right) \frac{1}{\sqrt{2\pi}} e^{ikx}$$

and now simplify to get

$$\frac{1}{3} \pi^2 + \sum_{k=1}^3 (-1)^k \left( \frac{4}{k^2} \right) \cos kx$$

Then a graph of this along with the graph of $y = x^2$ is given below. In this graph, the dashed graph is of $y = x^2$ and the solid line is the graph of the above Fourier series approximation. If we had taken the partial sum up to $n$ much bigger, it would have been very hard to distinguish between the graph of the partial sum of the Fourier series and the graph of the function it is approximating. This is in contrast to approximation by Taylor series in which you only get approximation at a point of a function and its derivatives. These are very close near the point of interest but typically fail to approximate the function on the entire interval.

### 13.2 Riesz Representation Theorem, Adjoint Map

The next theorem is one of the most important results in the theory of inner product spaces. It is called the Riesz representation theorem.

**Theorem 13.2.1** Let $f \in \mathcal{L}(X, \mathbb{F})$ where $X$ is an inner product space of dimension $n$. Then there exists a unique $z \in X$ such that for all $x \in X$,

$$f(x) = (x, z).$$

**Proof:** First I will verify uniqueness. Suppose $z_j$ works for $j = 1, 2$. Then for all $x \in X$,

$$0 = f(x) - f(x) = (x, z_1 - z_2)$$

and so $z_1 = z_2$. 
It remains to verify existence. By Lemma 11.4.13, there exists an orthonormal basis, \( \{u_j\}_{j=1}^n \).
If there is such a \( z \), then you would need \( f(u_j) = (u_j, z) \) and so you would need \( f(u_j) = (z, u_j) \).
Also you must have \( z = \sum_i (z, u_j) u_j \). Therefore, define
\[
z = \sum_{j=1}^n f(u_j) u_j.
\]
Then using Lemma 13.1.1,
\[
(x, z) = (x, \sum_{j=1}^n f(u_j) u_j) = \sum_{j=1}^n f(u_j) (x, u_j)
\]
\[
= f(\sum_{j=1}^n (x, u_j) u_j) = f(x) . \quad \blacksquare
\]

**Corollary 13.2.2** Let \( A \in \mathcal{L}(X, Y) \) where \( X \) and \( Y \) are two inner product spaces of finite dimension. Then there exists a unique \( A^* \in \mathcal{L}(Y, X) \) such that
\[
(Ax, y)_Y = (x, A^* y)_X \quad (13.4)
\]
for all \( x \in X \) and \( y \in Y \). The following formula holds
\[
(\alpha A + \beta B)^* = \overline{\alpha} A^* + \overline{\beta} B^*
\]

**Proof:** Let \( f_y \in \mathcal{L}(X, \mathbb{F}) \) be defined as
\[
f_y(x) = (Ax, y)_Y .
\]
Then by the Riesz representation theorem, there exists a unique element of \( X, A^*(y) \) such that
\[
(Ax, y)_Y = (x, A^*(y))_X .
\]
It only remains to verify that \( A^* \) is linear. Let \( a \) and \( b \) be scalars. Then for all \( x \in X, \)
\[
(x, A^*(ay_1 + by_2))_X = (Ax, (ay_1 + by_2))_Y
\]
\[
= \overline{\alpha} (Ax, y_1) + \overline{\beta} (Ax, y_2) = \overline{\alpha} (x, A^*(y_1)) + \overline{\beta} (x, A^*(y_2)) = (x, aA^*(y_1) + bA^*(y_2)).
\]
Since this holds for every \( x \), it follows
\[
A^*(ay_1 + by_2) = aA^*(y_1) + bA^*(y_2)
\]
which shows \( A^* \) is linear as claimed.

Consider the last assertion that \( * \) is conjugate linear.
\[
(x, (\alpha A + \beta B)^* y) = ((\alpha A + \beta B)x, y)
\]
\[
= \alpha (Ax, y) + \beta (Bx, y) = \alpha (x, A^* y) + \beta (x, B^* y)
\]
\[
= (x, \overline{\alpha} A^* y) + (x, \overline{\beta} A^* y) = (x, (\overline{\alpha} A^* + \overline{\beta} A^*) y).
\]
Since \( x \) is arbitrary,
\[
(\alpha A + \beta B)^* y = (\overline{\alpha} A^* + \overline{\beta} A^*) y
\]
and since this is true for all \( y, \)
\[
(\alpha A + \beta B)^* = \overline{\alpha} A^* + \overline{\beta} A^* . \quad \blacksquare
\]
**Definition 13.2.3** The linear map, $A^*$ is called the adjoint of $A$. In the case when $A: X \rightarrow X$ and $A = A^*$, $A$ is called a self adjoint map. Such a map is also called Hermitian.

**Theorem 13.2.4** Let $M$ be an $m \times n$ matrix. Then $M^* = (\overline{M})^T$ in words, the transpose of the conjugate of $M$ is equal to the adjoint.

**Proof:** Using the definition of the inner product in $\mathbb{C}^n$,

$$ (Mx, y) = (x, M^*y) \equiv \sum_i x_i \sum_j (M^*)_{ij} y_j = \sum_{i,j} (M^*)_{ij} \overline{y_j} x_i. $$

Also

$$ (Mx, y) = \sum_j \sum_i M_{ji} \overline{y_j} x_i. $$

Since $x, y$ are arbitrary vectors, it follows that $M_{ji} = (M^*)_{ij}$ and so, taking conjugates of both sides,

$$ M^*_{ij} = \overline{M}_{ji}. $$

The next theorem is interesting. You have a $p$ dimensional subspace of $\mathbb{F}^n$ where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. Of course this might be “slanted”. However, there is a linear transformation $Q$ which preserves distances which maps this subspace to $\mathbb{F}^p$.

**Theorem 13.2.5** Suppose $V$ is a subspace of $\mathbb{F}^n$ having dimension $p \leq n$. Then there exists a $Q \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ such that

$$ QV \subseteq \text{span} \{e_1, \ldots, e_p\} $$

and $|Qx| = |x|$ for all $x$. Also

$$ Q^*Q = QQ^* = I. $$

**Proof:** By Lemma 13.2.3 there exists an orthonormal basis for $V$, $\{v_i\}_{i=1}^p$. By using the Gram Schmidt process this may be extended to an orthonormal basis of the whole space $\mathbb{F}^n$,

$$ \{v_1, \ldots, v_p, v_{p+1}, \ldots, v_n\}. $$

Now define $Q \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ by $Q(v_i) \equiv e_i$ and extend linearly. If $\sum_{i=1}^n x_i v_i$ is an arbitrary element of $\mathbb{F}^n$,

$$ \left| Q \left( \sum_{i=1}^n x_i v_i \right) \right|^2 = \sum_{i=1}^n x_i e_i^2 = \sum_{i=1}^n |x_i|^2 = \sum_{i=1}^n x_i v_i^2. $$

It remains to verify that $Q^*Q = QQ^* = I$. To do so, let $x, y \in \mathbb{F}^n$. Then let $\omega$ be a complex number such that $|\omega| = 1$, $\omega (x, Q^*Qy - y) = |(x, Q^*Qy - y)|$. From what was just shown,

$$ (Q(\omega x + y), Q(\omega x + y)) = (\omega x + y, \omega x + y). $$

Thus

$$ |Qx|^2 + |Qy|^2 + 2 \Re \omega (Qx, Qy) = |x|^2 + |y|^2 + 2 \Re \omega (x, y) $$

and since $Q$ preserves norms, it follows that for all $x, y \in \mathbb{F}^n$,

$$ \Re \omega (Qx, Qy) = \Re \omega (x, Q^*Qy) = \omega \Re (x, y). $$

Thus

$$ 0 = \Re \omega ((x, Q^*Qy) - (x, y)) = \Re \omega (x, Q^*Qy - y) = |(x, Q^*Qy - y)| $$

for all $x, y$. Letting $x = Q^*Qy - y$, it follows $Q^*Qy = y$. Similarly $QQ^* = I$. 

**Definition 13.2.6** If $U \in \mathcal{L}(X, X)$ for $X$ an inner product space, then $U$ is called unitary if $U^*U = UU^* = I$. 

Note that it is actually shown that \( QV = \text{span}(e_1, \ldots, e_p) \) and that in case \( p = n \) one obtains that a linear transformation which maps an orthonormal basis to an orthonormal basis is unitary. Unitary matrices are also characterized by preserving length. More generally

**Lemma 13.2.7** Suppose \( R \in \mathcal{L}(X,Y) \) where \( X,Y \) are inner product spaces and \( R \) preserves distances. Then \( R^*R = I \).

**Proof:** Since \( R \) preserves distances, \( |Ru| = |u| \) for every \( u \). Let \( u, v \) be arbitrary vectors in \( X \) and let \( \theta \in C, \ |\theta| = 1, \) and \( \theta(R^*Ru - u, v) = |(R^*Ru - u, v)| \). Therefore from the axioms of the inner product,
\[
|u|^2 + |v|^2 + 2 \Re \theta(u, v) = |\theta u|^2 + |v|^2 + \theta(u, v) + \bar{\theta}(v, u) = |\theta u + v|^2 = (R(\theta u + v), R(\theta u + v))
\]
\[
= (R\theta u, R\theta u) + (Rv, Rv) + (R\theta u, Rv) + (Rv, R\theta u)
\]
\[
= |\theta u|^2 + |v|^2 + \theta(R^*Ru, v) + \bar{\theta}(v, R^*Ru)
\]
\[
= |u|^2 + |v|^2 + 2 \Re \theta(R^*Ru, v)
\]
and so for all \( u, v \),
\[
2 \Re \theta(R^*Ru - u, v) = 2 |(R^*Ru - u, v)| = 0
\]
Now let \( v = R^*Ru - u \). It follows that \( R^*Ru - u = 0 \). ■

**Corollary 13.2.8** Suppose \( U \in \mathcal{L}(X,X) \) where \( X \) is an inner product space. Then \( U \) is unitary if and only if \( |Ux| = |x| \) for all \( x \) so it preserves distance.

**Proof:** \( \Rightarrow \) If \( U \) is unitary, then \( |Ux|^2 = (Ux, Ux) = (U^*Ux, x) = (x, x) = |x|^2 \).
\( \Leftarrow \) If \( |Ux| = |x| \) for all \( x \) then by Lemma 13.2.7, \( U^*U = I \). Thus also \( UU^* = I \) because these are square matrices. Hence \( U \) is unitary. ■

Now here is an important result on factorization of an \( m \times n \) matrix. It is called a QR factorization.

**Theorem 13.2.9** Let \( A \) be an \( m \times n \) complex matrix. Then there exists a unitary \( Q \) and upper triangular \( R \) such that \( A = QR \).

**Proof:** This is obvious if \( m = 1 \). Suppose true for \( m - 1 \) and let
\[
A = \begin{pmatrix} a_1 & \cdots & a_n & a_{n+1} \end{pmatrix}
\]
Let \( Q_1 \) be a unitary matrix such that \( Q_1a_1 = |a_1|e_1 \) in case \( a_1 \neq 0 \). If \( a_1 = 0 \), let \( Q_1 = I \). Thus
\[
Q_1A = \begin{pmatrix} a & b \\ 0 & A_1 \end{pmatrix}
\]
where \( A_1 \) is \((m-1) \times (n-1)\). By induction, there exists \( Q_2' \) an \((m-1) \times (n-1)\) unitary matrix such that \( Q_2'A_1 = R' \), an upper triangular matrix. Then
\[
\begin{pmatrix} 1 & 0 \\ 0 & Q_2' \end{pmatrix} Q_1A = \begin{pmatrix} a & b \\ 0 & R' \end{pmatrix} = R
\]
Since the product of unitary matrices is unitary, there exists \( Q \) unitary such that \( Q^*A = R \) and so \( A = QR \). ■
13.3 Least Squares

A common problem in experimental work is to find a straight line which approximates as well as possible a collection of points in the plane \( \{(x_i, y_i)\}_{i=1}^n \). The usual way of dealing with these problems is by the method of least squares and it turns out that all these sorts of approximation problems can be reduced to \( Ax = b \) where the problem is to find the best \( x \) for solving this equation even when there is no solution.

**Lemma 13.3.1** Let \( V \) and \( W \) be finite dimensional inner product spaces and let \( A : V \to W \) be linear. For each \( y \in W \) there exists \( x \in V \) such that

\[
|Ax - y| \leq |Ax_1 - y|
\]

for all \( x_1 \in V \). Also, \( x \in V \) is a solution to this minimization problem if and only if \( x \) is a solution to the equation, \( A^*Ax = A^*y \).

**Proof:** By Theorem 13.1 on Page 208 there exists a point, \( Ax_0 \), in the finite dimensional subspace, \( A(V) \), of \( W \) such that for all \( x \in V, |Ax - y| \geq |Ax_0 - y| \). Also, from this theorem, this happens if and only if \( Ax_0 - y \) is perpendicular to every \( Ax \in A(V) \). Therefore, the solution is characterized by \( (Ax_0 - y, Ax) = 0 \) for all \( x \in V \) which is the same as saying \( (A^*Ax_0 - A^*y, x) = 0 \) for all \( x \in V \). In other words the solution is obtained by solving \( A^*Ax_0 = A^*y \) for \( x_0 \).

Consider the problem of finding the least squares regression line in statistics. Suppose you have given points in the plane, \( \{(x_i, y_i)\}_{i=1}^n \) and you would like to find constants \( m \) and \( b \) such that the line \( y = mx + b \) goes through all these points. Of course this will be impossible in general. Therefore, try to find \( m, b \) such that you do the best you can to solve the system

\[
\begin{pmatrix}
  y_1 \\
  \vdots \\
  y_n
\end{pmatrix} = \begin{pmatrix}
  x_1 & 1 \\
  \vdots & \vdots \\
  x_n & 1
\end{pmatrix} \begin{pmatrix}
  m \\
  b
\end{pmatrix}
\]

which is of the form \( y = Ax \). In other words try to make

\[
|A \begin{pmatrix}
  m \\
  b
\end{pmatrix} - \begin{pmatrix}
  y_1 \\
  \vdots \\
  y_n
\end{pmatrix}|^2
\]

as small as possible.

According to what was just shown, it is desired to solve the following for \( m \) and \( b \).

\[
A^*A \begin{pmatrix}
  m \\
  b
\end{pmatrix} = A^* \begin{pmatrix}
  y_1 \\
  \vdots \\
  y_n
\end{pmatrix}.
\]

Since \( A^* = A^T \) in this case,

\[
\begin{pmatrix}
  \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\
  \sum_{i=1}^n x_i & n
\end{pmatrix} \begin{pmatrix}
  m \\
  b
\end{pmatrix} = \begin{pmatrix}
  \sum_{i=1}^n x_i y_i \\
  \sum_{i=1}^n y_i
\end{pmatrix}
\]

Solving this system of equations for \( m \) and \( b \),

\[
m = -\frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i) + (\sum_{i=1}^n x_i y_i)n}{(\sum_{i=1}^n x_i^2)n - (\sum_{i=1}^n x_i)^2}
\]

and

\[
b = -\frac{(\sum_{i=1}^n x_i)\sum_{i=1}^n x_i y_i + (\sum_{i=1}^n y_i)\sum_{i=1}^n x_i^2}{(\sum_{i=1}^n x_i^2)n - (\sum_{i=1}^n x_i)^2}.
\]
One could clearly do a least squares fit for curves of the form \( y = ax^2 + bx + c \) in the same way. In this case you solve as well as possible for \( a, b, \) and \( c \) the system:

\[
\begin{pmatrix}
  x_1^2 & x_1 & 1 \\
  \vdots & \vdots & \vdots \\
  x_n^2 & x_n & 1 \\
\end{pmatrix}
\begin{pmatrix}
  a \\
  b \\
  c \\
\end{pmatrix}
= \begin{pmatrix}
  y_1 \\
  \vdots \\
  y_n \\
\end{pmatrix}
\]

using the same techniques.

### 13.4 Fredholm Alternative

The best context in which to study the Fredholm alternative is in inner product spaces. This is done here.

**Definition 13.4.1** Let \( S \) be a subset of an inner product space, \( X \). Define

\[
S^\perp \equiv \{ x \in X : (x, s) = 0 \text{ for all } s \in S \}.
\]

The following theorem also follows from the above lemma. It is sometimes called the Fredholm alternative.

**Theorem 13.4.2** Let \( A : V \rightarrow W \) where \( A \) is linear and \( V \) and \( W \) are inner product spaces. Then \( A(V) = \ker (A^*)^\perp \).

**Proof:** Let \( y = Ax \) so \( y \in A(V) \). Then if \( A^*z = 0 \),

\[
(y, z) = (Ax, z) = (x, A^*z) = 0
\]

showing that \( y \in \ker (A^*)^\perp \). Thus \( A(V) \subseteq \ker (A^*)^\perp \).

Now suppose \( y \in \ker (A^*)^\perp \). Does there exists \( x \) such that \( Ax = y \)? Since this might not be immediately clear, take the least squares solution to the problem. Thus let \( x \) be a solution to \( A^*Ax = A^*y \). It follows \( A^* (y - Ax) = 0 \) and so \( y - Ax \in \ker (A^*) \) which implies from the assumption about \( y \) that \( (y - Ax, y) = 0 \). Also, since \( Ax \) is the closest point to \( y \) in \( A(V) \), Theorem \[3.4.1\] on Page \[293\] implies that \( (y - Ax, Ax_1) = 0 \) for all \( x_1 \in V \). In particular this is true for \( x_1 = x \) and so

\[
0 = (y - Ax, y) - (y - Ax, Ax) = |y - Ax|^2,
\]

showing that \( y = Ax \). Thus \( A(V) \supseteq \ker (A^*)^\perp \). \( \square \)

**Corollary 13.4.3** Let \( A, V, \) and \( W \) be as described above. If the only solution to \( A^*y = 0 \) is \( y = 0 \), then \( A \) is onto \( W \).

**Proof:** If the only solution to \( A^*y = 0 \) is \( y = 0 \), then \( \ker (A^*) = \{0\} \) and so every vector from \( W \) is contained in \( \ker (A^*)^\perp \) and by the above theorem, this shows \( A(V) = W \). \( \square \)

### 13.5 The Determinant And Volume

The determinant is the essential algebraic tool which provides a way to give a unified treatment of the concept of \( p \) dimensional volume of a parallelepiped in \( \mathbb{R}^M \). Here is the definition of what is meant by such a thing.

**Definition 13.5.1** Let \( \mathbf{u}_1, \ldots, \mathbf{u}_p \) be vectors in \( \mathbb{R}^M, M \geq p \). The parallelepiped determined by these vectors will be denoted by \( P(\mathbf{u}_1, \ldots, \mathbf{u}_p) \) and it is defined as

\[
P(\mathbf{u}_1, \ldots, \mathbf{u}_p) \equiv \left\{ \sum_{j=1}^{p} s_j \mathbf{u}_j : s_j \in [0,1] \right\}.
\]
The volume of this parallelepiped is defined as

\[
\text{volume of } P(u_1, \ldots, u_p) \equiv v(P(u_1, \ldots, u_p)) \equiv (\det(G))^{1/2}.
\]

where \( G_{ij} = u_i \cdot u_j \).

If the vectors are dependent, this definition will give the volume to be 0.

First let's observe the last assertion is true. Say \( u_i = \sum_{j \neq i}^p \alpha_j u_j \). Then the \( i \)th row is a linear combination of the other rows and so from the properties of the determinant, the determinant of this matrix is indeed zero as it should be.

A parallelepiped is a sort of a squashed box. Here is a picture which shows the relationship between \( P(u_1, \ldots, u_{p-1}) \) and \( P(u_1, \ldots, u_p) \). In a sense, we can define the volume any way we want but if it is to be reasonable, the following relationship must hold. The appropriate definition of the volume of \( P(u_1, \ldots, u_p) \) in terms of \( P(u_1, \ldots, u_{p-1}) \) is

\[
v(P(u_1, \ldots, u_p)) = |u_p| \cos(\theta) v(P(u_1, \ldots, u_{p-1}))
\]

(13.5)

In the case where \( p = 1 \), the parallelepiped \( P(v) \) consists of the single vector and the one dimensional volume should be \( |v| = (v^Tv)^{1/2} = (v \cdot v)^{1/2} \). Now having made this definition, I will show that this is the appropriate definition of \( p \) dimensional volume for every \( p \).

**Definition 13.5.2** Let \( \{u_1, \ldots, u_p\} \) be vectors in \( \mathbb{R}^M \). Then letting \( U = \begin{pmatrix} u_1 & u_2 & \cdots & u_p \end{pmatrix} \),

\[
v(P(u_1, \ldots, u_p)) \equiv \det(U^T U)^{1/2} = \det(G)^{1/2}, \ G_{ij} = u_i \cdot u_j
\]

As just pointed out, this is the only reasonable definition of volume in the case of one vector. The next theorem shows that it is the only reasonable definition of volume of a parallelepiped in the case of \( p \) vectors because (13.7) holds.

**Theorem 13.5.3** With the above definition of volume, (13.7) holds.

**Proof:** To check whether this is so, it is necessary to find \(|\cos(\theta)|\). This involves finding a vector perpendicular to \( P(u_1, \ldots, u_{p-1}) \). Let \( \{w_1, \ldots, w_p\} \) be an orthonormal basis for span \( (u_1, \ldots, u_p) \) such that span \( (w_1, \ldots, w_k) = \text{span} (u_1, \ldots, u_k) \) for each \( k \leq p \). Such an orthonormal basis exists because of the Gram-Schmidt procedure. First note that since \( \{w_k\}_{k=1}^p \) is an orthonormal basis for span \( (u_1, \ldots, u_p) \),

\[
\begin{align*}
 u_j &= \sum_{k=1}^p (u_j \cdot w_k) w_k, \\
 u_j \cdot u_i &= \sum_{k=1}^p (u_j \cdot w_k) (u_i \cdot w_k)
\end{align*}
\]

Therefore, for each \( k \leq p \), the \( ij \)th entry of \( U^T U \) is just

\[
\begin{align*}
 u_j^T u_j &= u_j \cdot u_j = \sum_{r=1}^p (u_i \cdot w_r) (w_r \cdot u_j)
\end{align*}
\]

Thus this matrix is the product of the two \( p \times p \) matrices, \( M_{ij} = u_i \cdot w_j \) and its transpose \( (M^T)_{ij} = w_j \cdot u_i \).

It follows

\[
\det(U^T U) = \det(G)
\]
Thus if $\mathbf{v}(\mathbf{u}_1, \cdots, \mathbf{u}_p) = \pm A_{1k}$ where $A_{1k}$ is the $1k^{th}$ cofactor of the above matrix, equal to $\mathbf{N} \cdot \mathbf{w}_k$. Now it follows from the formula for expansion of a determinant along the top row that for each $j \leq p$,

$$\mathbf{N} \cdot \mathbf{u}_j = \mathbf{N} \sum_{k=1}^{p} (\mathbf{u}_j \cdot \mathbf{w}_k) \mathbf{w}_k = \sum_{k=1}^{p} (\mathbf{u}_j \cdot \mathbf{w}_k) \mathbf{N} \cdot \mathbf{w}_k = \sum_{k=1}^{p} (\mathbf{u}_j \cdot \mathbf{w}_k) A_{1k}$$

Thus if $j \leq p - 1$

$$\mathbf{N} \cdot \mathbf{u}_j = \det \begin{pmatrix} (\mathbf{u}_j \cdot \mathbf{w}_1) & (\mathbf{u}_j \cdot \mathbf{w}_2) & \cdots & (\mathbf{u}_j \cdot \mathbf{w}_p) \\ (\mathbf{u}_1 \cdot \mathbf{w}_1) & (\mathbf{u}_1 \cdot \mathbf{w}_2) & \cdots & (\mathbf{u}_1 \cdot \mathbf{w}_p) \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{u}_{p-1} \cdot \mathbf{w}_1) & (\mathbf{u}_{p-1} \cdot \mathbf{w}_2) & \cdots & (\mathbf{u}_{p-1} \cdot \mathbf{w}_p) \end{pmatrix} = 0$$

because the matrix has two equal rows. Therefore, $\mathbf{N}$ points in the direction of the normal vector in the above picture or else it points in the opposite direction to this vector. If $j = p$, $\mathbf{N} \cdot \mathbf{u}_p$ equals $\pm v(\mathbf{P} (\mathbf{u}_1, \cdots, \mathbf{u}_p))$. Thus $|\mathbf{N} \cdot \mathbf{u}_p| = v(\mathbf{P} (\mathbf{u}_1, \cdots, \mathbf{u}_p))$. From the geometric description of the dot product,

$$v(\mathbf{P} (\mathbf{u}_1, \cdots, \mathbf{u}_p)) = |\mathbf{N}| |\mathbf{u}_p| \cos(\theta)$$

Now $\mathbf{N}$ is perpendicular to each $\mathbf{w}_k$ and $\mathbf{u}_k$ for $k \leq p - 1$ and so

$$\mathbf{N} = (\mathbf{N} \cdot \mathbf{w}_p) \mathbf{w}_p = A_{1p} \mathbf{w}_p = \pm v(\mathbf{P} (\mathbf{u}_1, \cdots, \mathbf{u}_{p-1})) \mathbf{w}_p$$

Thus $|\mathbf{N}| = v(\mathbf{P} (\mathbf{u}_1, \cdots, \mathbf{u}_{p-1}))$ and so

$$v(\mathbf{P} (\mathbf{u}_1, \cdots, \mathbf{u}_p)) = v(\mathbf{P} (\mathbf{u}_1, \cdots, \mathbf{u}_{p-1})) |\mathbf{u}_p| \cos(\theta) \Box$$

The theorem shows that the only reasonable definition of $p$ dimensional volume of a parallelepiped is the one given in the above definition. Recall that these vectors are in $\mathbb{R}^M$. What is the role of $\mathbb{R}^M$? It is just to provide an inner product. That is its only function.
13.6 Exercises

1. Find the best solution to the system
   \[ \begin{align*}
   x + 2y &= 6 \\
   2x - y &= 5 \\
   3x + 2y &= 0
   \end{align*} \]

2. Find an orthonormal basis for \( \mathbb{R}^3 \), \( \{w_1, w_2, w_3\} \) given that \( w_1 \) is a multiple of the vector \( (1, 1, 2) \).

3. Suppose \( A = A^T \) is a symmetric real \( n \times n \) matrix which has all positive eigenvalues. Define
   \[(x, y) \equiv (Ax, y).\]
   Show this is an inner product on \( \mathbb{R}^n \). What does the Cauchy Schwarz inequality say in this case?

4. Let \( ||x||_\infty \equiv \max \{|x_j| : j = 1, 2, \ldots, n\} \). Show this is a norm on \( \mathbb{C}^n \). Here
   \[ x = \left( \begin{array}{c} x_1 \\
   \vdots \\
   x_n \end{array} \right)^T. \]
   Show
   \[ ||x||_\infty \leq |x| \equiv (x, x)^{1/2} \]
   where the above is the usual inner product on \( \mathbb{C}^n \).

5. Let \( ||x||_1 \equiv \sum_{j=1}^n |x_j| \). Show this is a norm on \( \mathbb{C}^n \). Here
   \[ x = \left( \begin{array}{c} x_1 \\
   \vdots \\
   x_n \end{array} \right)^T. \]
   Show
   \[ ||x||_1 \geq |x| \equiv (x, x)^{1/2} \]
   where the above is the usual inner product on \( \mathbb{C}^n \). Show there cannot exist an inner product such that this norm comes from the inner product as described above for inner product spaces.

6. Show that if \( ||\cdot|| \) is any norm on any vector space, then \( ||x|| - ||y|| \leq ||x - y|| \).

7. Relax the assumptions in the axioms for the inner product. Change the axiom about \( (x, x) \geq 0 \) and equals 0 if and only if \( x = 0 \) to simply read \( (x, x) \geq 0 \). Show the Cauchy Schwarz inequality still holds in the following form.
   \[ |(x, y)| \leq (x, x)^{1/2} (y, y)^{1/2}. \]

8. Let \( H \) be an inner product space and let \( \{u_k\}_{k=1}^n \) be an orthonormal basis for \( H \). Show
   \[ (x, y) = \sum_{k=1}^n (x, u_k) \overline{(y, u_k)}. \]

9. Let the vector space \( V \) consist of real polynomials of degree no larger than 3. Thus a typical vector is a polynomial of the form \( a + bx + cx^2 + dx^3 \). For \( p, q \in V \) define the inner product,
   \[ (p, q) \equiv \int_0^1 p(x) q(x) \, dx. \]
   Show this is indeed an inner product. Then state the Cauchy Schwarz inequality in terms of this inner product. Show \( \{1, x, x^2, x^3\} \) is a basis for \( V \). Finally, find an orthonormal basis for \( V \). This is an example of some orthonormal polynomials.

10. Let \( P_n \) denote the polynomials of degree no larger than \( n - 1 \) which are defined on an interval \( [a, b] \). Let \( \{x_1, \cdots, x_n\} \) be \( n \) distinct points in \( [a, b] \). Now define for \( p, q \in P_n, \)
    \[ (p, q) \equiv \sum_{j=1}^n p(x_j) \overline{q(x_j)}. \]
    Show this yields an inner product on \( P_n \). **Hint:** Most of the axioms are obvious. The one which says \( (p, p) = 0 \) if and only if \( p = 0 \) is the only interesting one. To verify this one, note that a nonzero polynomial of degree no more than \( n - 1 \) has at most \( n - 1 \) zeros.
11. Let $C([0,1])$ denote the vector space of continuous real valued functions defined on $[0,1]$. Let the inner product be given as

$$(f,g) = \int_0^1 f(x)g(x)\,dx$$

Show this is an inner product. Also let $V$ be the subspace described in Problem 10. Using the result of this problem, find the vector in $V$ which is closest to $x^4$.

12. A regular Sturm-Liouville problem involves the differential equation, for an unknown function of $x$ which is denoted here by $y$,

$$(p(x)y')' + (\lambda q(x) + r(x))y = 0, \quad x \in [a,b]$$

and it is assumed that $p(t), q(t) > 0$ for any $t \in [a,b]$ and also there are boundary conditions,

$$C_1 y(a) + C_2 y'(a) = 0$$
$$C_3 y(b) + C_4 y'(b) = 0$$

where

$$C_1^2 + C_2^2 > 0, \quad C_3^2 + C_4^2 > 0.$$ 

There is an immense theory connected to these important problems. The constant, $\lambda$ is called an eigenvalue. Show that if $y$ is a solution to the above problem corresponding to $\lambda = \lambda_2 \neq \lambda_1$ and if $z$ is a solution corresponding to $\lambda = \lambda_2 \neq \lambda_1$, then

$$\int_a^b q(x)y(x)z(x)\,dx = 0.$$ 

and this defines an inner product. **Hint:** Do something like this:

$$(p(x)y')'z + (\lambda_1 q(x) + r(x))yz = 0,$$

$$(p(x)z')'y + (\lambda_2 q(x) + r(x))zy = 0.$$ 

Now subtract and either use integration by parts or show

$$(p(x)y')'z - (p(x)z')'y = ((p(x)y')z - (p(x)z')y)'$$

and then integrate. Use the boundary conditions to show that $y'(a)z(a) - z'(a)y(a) = 0$ and $y'(b)z(b) - z'(b)y(b) = 0$. The formula, \(13.6\) is called an orthogonality relation. It turns out there are typically infinitely many eigenvalues and it is interesting to write given functions as an infinite series of these “eigenfunctions”.

13. Consider the continuous functions defined on $[0,\pi]$, $C([0,\pi])$. Show $(f,g) = \int_0^\pi f\,g\,dx$ is an inner product on this vector space. Show the functions $\left\{\sqrt{\frac{2}{n}} \sin(nx)\right\}_{n=1}^\infty$ are an orthonormal set. What does this mean about the dimension of the vector space $C([0,\pi])$? Now let $V_N = \text{span}\left(\sqrt{\frac{2}{\pi}} \sin(x), \cdots, \sqrt{\frac{2}{\pi}} \sin(Nx)\right)$. For $f \in C([0,\pi])$ find a formula for the vector in $V_N$ which is closest to $f$ with respect to the norm determined from the above inner product. This is called the $N$th partial sum of the Fourier series of $f$. An important problem is to determine whether and in what way this Fourier series converges to the function $f$. The norm which comes from this inner product is sometimes called the mean square norm.

14. Consider the subspace $V \equiv \ker(A)$ where

$$A = \begin{pmatrix}
1 & 4 & -1 & -1 \\
2 & 1 & 2 & 3 \\
4 & 9 & 0 & 1 \\
5 & 6 & 3 & 4
\end{pmatrix}$$

Find an orthonormal basis for $V$. **Hint:** You might first find a basis and then use the Gram Schmidt procedure.
13.6. EXERCISES

15. The Gram Schmidt process starts with a basis for a subspace \( \{v_1, \ldots, v_n\} \) and produces an orthonormal basis for the same subspace \( \{u_1, \ldots, u_n\} \) such that

\[
\text{span}(v_1, \ldots, v_k) = \text{span}(u_1, \ldots, u_k)
\]

for each \( k \). Show that in the case of \( \mathbb{R}^m \) the QR factorization does the same thing. More specifically, if

\[
A = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix}
\]

and if

\[
A = QR = \begin{pmatrix} q_1 & \cdots & q_n \end{pmatrix} R
\]

then the vectors \( \{q_1, \ldots, q_n\} \) is an orthonormal set of vectors and for each \( k \),

\[
\text{span}(q_1, \ldots, q_k) = \text{span}(v_1, \ldots, v_k)
\]

16. Verify the parallelogram identify for any inner product space,

\[
|x + y|^2 + |x - y|^2 = 2|x|^2 + 2|y|^2.
\]

Why is it called the parallelogram identity?

17. Let \( H \) be an inner product space and let \( K \subseteq H \) be a nonempty convex subset. This means that if \( k_1, k_2 \in K \), then the line segment consisting of points of the form

\[
tk_1 + (1 - t)k_2 \quad \text{for} \quad t \in [0, 1]
\]

is also contained in \( K \). Suppose for each \( x \in H \), there exists \( Px \) defined to be a point of \( K \) closest to \( x \). Show that \( Px \) is unique so that \( P \) actually is a map. **Hint:** Suppose \( z_1 \) and \( z_2 \) both work as closest points. Consider the midpoint, \( (z_1 + z_2)/2 \) and use the parallelogram identity of Problem 16 in an auspicious manner.

18. In the situation of Problem 17 suppose \( K \) is a closed convex subset and that \( H \) is complete. This means if every Cauchy sequence converges. Recall a sequence \( \{k_n\} \) is a Cauchy sequence if for every \( \varepsilon > 0 \), there exists \( N_\varepsilon \) such that whenever \( m, n > N_\varepsilon \), it follows \( |k_m - k_n| < \varepsilon \). Let \( \{k_n\} \) be a sequence of points of \( K \) such that

\[
\lim_{n \to \infty} |x - k_n| = \inf \{ |x - k| : k \in K \}
\]

This is called a minimizing sequence. Show there exists a unique \( k \in K \) such that

\[
\lim_{n \to \infty} |k_n - k|
\]

and that \( k = Px \). That is, there exists a well defined projection map onto the convex subset of \( H \). **Hint:** Use the parallelogram identity in an auspicious manner to show \( \{k_n\} \) is a Cauchy sequence which must therefore converge. Since \( K \) is closed it follows this will converge to something in \( K \) which is the desired vector.

19. Let \( H \) be an inner product space which is also complete and let \( P \) denote the projection map onto a convex closed subset, \( K \). Show this projection map is characterized by the inequality

\[
\text{Re} \left( k - Px, x - Px \right) \leq 0
\]

for all \( k \in K \). That is, a point \( z \in K \) equals \( Px \) if and only if the above variational inequality holds. This is what that inequality is called. This is because \( k \) is allowed to vary and the inequality continues to hold for all \( k \in K \).
20. Using Problem 23 and Problems 24 - 25 show the projection map, $P$ onto a closed convex subset is Lipschitz continuous with Lipschitz constant 1. That is

$$|Px - Py| \leq |x - y|$$

21. Give an example of two vectors in $\mathbb{R}^4$ or $\mathbb{R}^3$ $x, y$ and a subspace $V$ such that $x \cdot y = 0$ but $Px \cdot Py \neq 0$ where $P$ denotes the projection map which sends $x$ to its closest point on $V$.

22. Suppose you are given the data, $(1, 2), (2, 4), (3, 8), (0, 0)$. Find the linear regression line using the formulas derived above. Then graph the given data along with your regression line.

23. Generalize the least squares procedure to the situation in which data is given and you desire to find $a, b, c$ in order to minimize the error. Could this be generalized to higher dimensions? How about more functions?

24. Let $A \in \mathcal{L}(X, Y)$ where $X$ and $Y$ are finite dimensional vector spaces with the dimension of $X$ equal to $n$. Define $\text{rank}(A) \equiv \dim(A(X))$ and $\text{nullity}(A) \equiv \dim(\ker(A))$. Show that $\text{nullity}(A) + \text{rank}(A) = \dim(X)$. \textbf{Hint}: Let $\{x_i\}_{i=1}^r$ be a basis for $\ker(A)$ and let $\{x_i\}_{i=1}^r \cup \{y_i\}_{i=1}^{n-r}$ be a basis for $X$. Then show that $\{Ay_i\}_{i=1}^{n-r}$ is linearly independent and spans $AX$.

25. Let $A$ be an $m \times n$ matrix. Show the column rank of $A$ equals the column rank of $A^*A$. Next verify column rank of $A^*A$ is no larger than column rank of $A^*$. Next justify the following inequality to conclude the column rank of $A$ equals the column rank of $A^*$.

$$\text{rank}(A) = \text{rank}(A^*A) \leq \text{rank}(A^*) \leq \text{rank}(AA^*) \leq \text{rank}(A)$$

\textbf{Hint}: Start with an orthonormal basis, $\{Ax_j\}_{j=1}^r$ of $A(\mathbb{F}^n)$ and verify $\{A^*Ax_j\}_{j=1}^r$ is a basis for $A^*A(\mathbb{F}^n)$.

26. Let $A$ be a real $m \times n$ matrix and let $A = QR$ be the QR factorization with $Q$ orthogonal and $R$ upper triangular. Show that there exists a solution $x$ to the equation

$$R^T Rx = R^T Q^T b$$

and that this solution is also a least squares solution defined above such that $A^T A x = A^T b$.

27. Here are three vectors in $\mathbb{R}^4: (1, 2, 0, 3)^T, (2, 1, -3, 2)^T, (0, 0, 1, 2)^T$. Find the three dimensional volume of the parallelepiped determined by these three vectors.

28. Here are two vectors in $\mathbb{R}^4: (1, 2, 0, 3)^T, (2, 1, -3, 2)^T$. Find the volume of the parallelepiped determined by these two vectors.

29. Here are three vectors in $\mathbb{R}^2: (1, 2)^T, (2, 1)^T, (0, 1)^T$. Find the three dimensional volume of the parallelepiped determined by these three vectors. Recall that from the above theorem, this should equal 0.

30. Find the equation of the plane through the three points $(1, 2, 3), (2, -3, 1), (1, 1, 7)$.

31. Let $T$ map a vector space $V$ to itself. Explain why $T$ is one to one if and only if $T$ is onto. It is in the text, but do it again in your own words.

32. Let all matrices be complex with complex field of scalars and let $A$ be an $n \times n$ matrix and $B$ a $m \times m$ matrix while $X$ will be an $n \times m$ matrix. The problem is to consider solutions to Sylvester’s equation. Solve the following equation for $X$

$$AX - XB = C$$
where $C$ is an arbitrary $n \times m$ matrix. Show there exists a unique solution if and only if $\sigma(A) \cap \sigma(B) = \emptyset$. **Hint:** If $q(\lambda)$ is a polynomial, show first that if $AX - XB = 0$, then $q(A)X - Xq(B) = 0$. Next define the linear map $T$ which maps the $n \times m$ matrices to the $n \times m$ matrices as follows.

$$TX = AX - XB$$

Show that the only solution to $TX = 0$ is $X = 0$ so that $T$ is one to one if and only if $\sigma(A) \cap \sigma(B) = \emptyset$. Do this by using the first part for $q(\lambda)$ the characteristic polynomial for $B$ and then use the Cayley Hamilton theorem. Explain why $q(A)^{-1}$ exists if and only if the condition $\sigma(A) \cap \sigma(B) = \emptyset$.

33. Compare Definition 13.5.2 with the Binet Cauchy theorem, Theorem 8.4.5. What is the geometric meaning of the Binet Cauchy theorem in this context?

34. Let $U, H$ be finite dimensional inner product spaces. (More generally, complete inner product spaces.) Let $A$ be a linear map from $U$ to $H$. Thus $AU$ is a subspace of $H$. For $g \in AU$, define $A^{-1}g$ to be the unique element of $\{x : Ax = g\}$ which is closest to $0$. Then define $(h, g)_{AU} = (A^{-1}g, A^{-1}h)_U$. Show that this is a well defined inner product. Let $U, H$ be finite dimensional inner product spaces. (More generally, complete inner product spaces.) Let $A$ be a linear map from $U$ to $H$. Thus $AU$ is a subspace of $H$. For $g \in AU$, define $A^{-1}g$ to be the unique element of $\{x : Ax = g\}$ which is closest to $0$. Then define $(h, g)_{AU} = (A^{-1}g, A^{-1}h)_U$. Show that this is a well defined inner product and that if $A$ is one to one, then $\|h\|_{AU} = \|A^{-1}h\|_U$ and $\|Ax\|_{AU} = \|x\|_U$. 
Chapter 14

Matrices And The Inner Product

14.1 Schur’s Theorem, Hermitian Matrices

Every matrix is related to an upper triangular matrix in a particularly significant way. This is Schur’s theorem and it is the most important theorem in the spectral theory of matrices. The important result which makes this theorem possible is the Gram Schmidt procedure of Lemma 11.4.13.

Definition 14.1.1 An \( n \times n \) matrix \( U \) is unitary if \( U^* U = I = U U^* \) where \( U^* \) is defined to be the transpose of the conjugate of \( U \). Thus \( U_{ij} = U_{ji}^* \). Note that every real orthogonal matrix is unitary.

Note that if \( U = (v_1 \cdots v_n) \) where the \( v_k \) are orthonormal vectors in \( \mathbb{C}^n \), then \( U \) is unitary. This follows because the \( ij \)th entry of \( U^* U \) is \( v_i^T v_j = \delta_{ij} \) since the \( v_i \) are assumed orthonormal.

Lemma 14.1.2 The following holds. \( (AB)^* = B^* A^* \).

Proof: From the definition and remembering the properties of complex conjugation,

\[
((AB)^*)_{ij} = (AB)_{ji} = \sum_k A_{ik} B_{kj} = \sum_k A_{ik} B_{kj}^*
\]

\[
= \sum_k B^*_{jk} A_{ki} = (B^* A^*)_{ji}
\]

Theorem 14.1.3 Let \( A \) be an \( n \times n \) matrix. Then there exists a unitary matrix \( U \) such that

\[
U^* AU = T,
\]

where \( T \) is an upper triangular matrix having the eigenvalues of \( A \) on the main diagonal listed according to multiplicity as roots of the characteristic equation. If \( A \) is a real matrix having all real eigenvalues, then \( U \) can be chosen to be an orthogonal real matrix.

Proof: The theorem is clearly true if \( A \) is a \( 1 \times 1 \) matrix. Just let \( U = 1 \), the \( 1 \times 1 \) matrix which has entry 1. Suppose it is true for \( (n-1) \times (n-1) \) matrices, \( n \geq 2 \) and let \( A \) be an \( n \times n \) matrix. Then let \( v_1 \) be a unit eigenvector for \( A \). Then there exists \( \lambda_1 \) such that

\[
Av_1 = \lambda_1 v_1, \quad |v_1| = 1.
\]

Extend \( \{v_1\} \) to a basis and then use the Gram - Schmidt process to obtain \( \{v_1, \cdots, v_n\} \), an orthonormal basis of \( \mathbb{C}^n \). Let \( U_0 \) be a matrix whose \( i \)th column is \( v_i \) so that \( U_0 \) is unitary. Consider \( U_0^* AU_0 \)

\[
U_0^* AU_0 = \begin{pmatrix} v_1^T \\ \vdots \\ v_n^T \end{pmatrix} \begin{pmatrix} Av_1 & \cdots & Av_n \end{pmatrix} = \begin{pmatrix} v_1^T \\ \vdots \\ v_n^T \end{pmatrix} \begin{pmatrix} \lambda_1 v_1 & \cdots & Av_n \end{pmatrix}
\]

309
Thus \(U_0^*AU_0\) is of the form
\[
\begin{pmatrix}
\lambda_1 & a \\
0 & A_1
\end{pmatrix}
\]
where \(A_1\) is an \((n-1) \times (n-1)\) matrix. Now by induction, there exists an \((n-1) \times (n-1)\) unitary matrix \(\tilde{U}_1\) such that \(\tilde{U}_1^*A_1\tilde{U}_1 = T_{n-1}\), an upper triangular matrix. Consider

\[
U_1 \equiv \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U}_1 \end{pmatrix}.
\]

Then
\[
U_1^*U_1 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U}_1^* \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U}_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & I_{n-1} \end{pmatrix}
\]

Also
\[
U_1^*U_0^*AU_0U_1 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U}_1^* \end{pmatrix} \begin{pmatrix} \lambda_1 & * \\ 0 & A_1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U}_1 \end{pmatrix} = \begin{pmatrix} \lambda_1 & * \\ 0 & T_{n-1} \end{pmatrix} \equiv T
\]

where \(T\) is upper triangular. Then let \(U = U_0U_1\). It is clear that this is unitary because both matrices preserve distance. Therefore, so does the product and hence \(U\). Alternatively,
\[
I = U_0U_1^*U_0^* = (U_0U_1)(U_0U_1)^*
\]
and so, it follows that \(A\) is similar to \(T\) and that \(U_0U_1\) is unitary. Hence \(A\) and \(T\) have the same characteristic polynomials, and since the eigenvalues of \(T\) \((A)\) are the diagonal entries listed with multiplicity, this proves the main conclusion of the theorem. In case \(A\) is real with all real eigenvalues, the above argument can be repeated word for word using only the real dot product to show that \(U\) can be taken to be real and orthogonal. ■

As a simple consequence of the above theorem, here is an interesting lemma.

**Lemma 14.1.4** Let \(A\) be of the form
\[
A = \begin{pmatrix}
P_1 & \cdots & * \\
\vdots & \ddots & \vdots \\
0 & \cdots & P_s
\end{pmatrix}
\]
where \(P_k\) is an \(m_k \times m_k\) matrix. Then
\[
\det (A) = \prod_k \det (P_k).
\]

**Proof:** Let \(U_k\) be an \(m_k \times m_k\) unitary matrix such that
\[
U_k^*P_kU_k = T_k
\]
where \(T_k\) is upper triangular. Then letting \(U\) denote the block diagonal matrix, having the \(U_i\) as the blocks on the diagonal,
\[
U = \begin{pmatrix} U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_s \end{pmatrix}, \quad U^* = \begin{pmatrix} U_1^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_s^* \end{pmatrix}
\]
and
\[
\begin{pmatrix}
U_1^* & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & U_s^*
\end{pmatrix}
\begin{pmatrix}
P_1 & \cdots & * \\
\vdots & \ddots & \vdots \\
0 & \cdots & P_s
\end{pmatrix}
\begin{pmatrix}
U_1 & \cdots & 0 \\
\ddots & \ddots & \ddots \\
0 & \cdots & U_s
\end{pmatrix} =
\begin{pmatrix}
T_1 & \cdots & * \\
\ddots & \ddots & \ddots \\
0 & \cdots & T_s
\end{pmatrix}
\]

and so
\[
\det (A) = \prod_k \det (T_k) = \prod_k \det (P_k). \quad \blacksquare
\]

**Definition 14.1.5** An \( n \times n \) matrix \( A \) is called **Hermitian** if \( A = A^* \). Thus a real symmetric \((A = A^T)\) matrix is Hermitian.

The following is the major result about Hermitian matrices. It says that any Hermitian matrix is unitarily similar because the matrix \( U \) in the following theorem which gives the similarity transformation is a unitary matrix.

**Theorem 14.1.6** If \( A \) is an \( n \times n \) Hermitian matrix, there exists a unitary matrix \( U \) such that
\[
U^*AU = D
\]
where \( D \) is a real diagonal matrix. That is, \( D \) has nonzero entries only on the main diagonal and these are real. Furthermore, the columns of \( U \) are an orthonormal basis of eigenvectors for \( \mathbb{C}^n \). If \( A \) is real and symmetric, then \( U \) can be assumed to be a real orthogonal matrix and the columns of \( U \) form an orthonormal basis for \( \mathbb{R}^n \).

**Proof:** From Schur’s theorem above, there exists \( U \) unitary (real and orthogonal if \( A \) is real) such that
\[
U^*AU = T
\]
where \( T \) is an upper triangular matrix. Then from Lemma 14.1.2
\[
T^* = (U^*AU)^* = U^*A^*U = U^*AU = T.
\]
Thus \( T = T^* \) and \( T \) is upper triangular. This can only happen if \( T \) is really a diagonal matrix having real entries on the main diagonal. (If \( i \neq j \), one of \( T_{ij} \) or \( T_{ji} \) equals zero. But \( T_{ij} = \overline{T_{ji}} \), and so they are both zero. Also \( T_{ii} = \overline{T_{ii}} \).)

Finally, let
\[
U = \begin{pmatrix}
u_1 & u_2 & \cdots & u_n
\end{pmatrix}
\]
where the \( u_i \) denote the columns of \( U \) and
\[
D = \begin{pmatrix}
\lambda_1 & 0 \\
& \ddots \\
0 & \lambda_n
\end{pmatrix}
\]

The equation, \( U^*AU = D \) implies
\[
AU = \begin{pmatrix}
Au_1 & Au_2 & \cdots & Au_n
\end{pmatrix}
= UD = \begin{pmatrix}
\lambda_1 u_1 & \lambda_2 u_2 & \cdots & \lambda_n u_n
\end{pmatrix}
\]
where the entries denote the columns of \( AU \) and \( UD \) respectively. Therefore, \( Au_i = \lambda_i u_i \) and since the matrix is unitary, the \( i^{th} \) entry of \( U^*U \) equals \( \delta_{ij} \) and so
\[
\delta_{ij} = u_i^T u_j = \overline{u_i} u_j = u_j \cdot u_i.
\]
This proves the corollary because it shows the vectors \( \{u_i\} \) form an orthonormal basis. In case \( A \) is real and symmetric, simply ignore all complex conjugations in the above argument. ■

This theorem is particularly nice because the diagonal entries are all real. What of a matrix which is unitarily similar to a diagonal matrix without assuming the diagonal entries are real? That is, \( A \) is an \( n \times n \) matrix with

\[
U^*AU = D
\]

Then this requires

\[
U^*A^*U = D^*
\]

and so since the two diagonal matrices commute,

\[
AA^* = UDU^*UD^*U^* = UDD^*U^* = UD^*DU^*
\]

\[
= UD^*U^*UDU^* = A^*A
\]

The following definition describes these matrices.

**Definition 14.1.7** An \( n \times n \) matrix is normal means: \( A^*A = AA^* \).

We just showed that if \( A \) is unitarily similar to a diagonal matrix, then it is normal. The converse is also true. This involves the following lemma.

**Lemma 14.1.8** If \( T \) is upper triangular and normal, then \( T \) is a diagonal matrix.

**Proof:** This is obviously true if \( T \) is \( 1 \times 1 \). In fact, it can’t help being diagonal in this case. Suppose then that the lemma is true for \((n-1) \times (n-1)\) matrices and let \( T \) be an upper triangular normal \( n \times n \) matrix. Thus \( T \) is of the form

\[
T = \begin{pmatrix}
t_{11} & a^* \\
0 & T_1
\end{pmatrix}, \quad T^* = \begin{pmatrix}
\overline{t_{11}} & 0^T \\
a & T_1^*
\end{pmatrix}
\]

Then

\[
TT^* = \begin{pmatrix}
t_{11} & a^* \\
0 & T_1
\end{pmatrix} \begin{pmatrix}
\overline{t_{11}} & 0^T \\
a & T_1^*
\end{pmatrix} = \begin{pmatrix}
|t_{11}|^2 + a^*a & a^*T_1^* \\
aT_1 & T_1T_1^*
\end{pmatrix}
\]

\[
T^*T = \begin{pmatrix}
\overline{t_{11}} & 0^T \\
a & T_1^*
\end{pmatrix} \begin{pmatrix}
t_{11} & a^* \\
0 & T_1
\end{pmatrix} = \begin{pmatrix}
|t_{11}|^2 & \overline{t_{11}}a^* \\
a^*a & a^* + T_1^*T_1
\end{pmatrix}
\]

Since these two matrices are equal, it follows \( a = 0 \). But now it follows that \( T_1^*T_1 = T_1T_1^* \) and so by induction \( T_1 \) is a diagonal matrix \( D_1 \). Therefore,

\[
T = \begin{pmatrix}
t_{11} & 0^T \\
0 & D_1
\end{pmatrix}
\]

a diagonal matrix. ■

**Theorem 14.1.9** An \( n \times n \) matrix is unitarily similar to a diagonal matrix if and only if it is normal.

**Proof:** It was already shown above that if \( A \) is similar to a diagonal matrix then it is normal. Suppose now that it is normal. By Schur’s theorem, there is a unitary matrix \( U \) such that

\[
U^*AU = T
\]

where \( T \) is upper triangular. In fact \( T \) is also normal. Therefore, from the above lemma, \( T \) is a diagonal matrix and this will complete the proof if it is shown that \( T \) is normal. However,

\[
T^*T = U^*A^*UU^*AU = U^*A^*AU
\]

\[
TT^* = U^*A^*UU^*A^*U = U^*AA^*U = U^*A^*AU
\]

showing that indeed \( T^*T = TT^* \). ■

In fact, if \( A \) is normal and \( B \) is unitarily similar to \( A \), then \( B \) will also be normal. This was what was shown in the second half of this theorem.
14.2 Quadratic Forms

Definition 14.2.1 A quadratic form in three dimensions is an expression of the form

\[
\begin{pmatrix} x & y & z \end{pmatrix} A \begin{pmatrix} x \\ y \\ z \end{pmatrix}
\]

(14.3)

where \( A \) is a \( 3 \times 3 \) symmetric matrix. In higher dimensions the idea is the same except you use a larger symmetric matrix in place of \( A \). In two dimensions \( A \) is a \( 2 \times 2 \) matrix.

For example, consider

\[
\begin{pmatrix} x & y & z \end{pmatrix} \begin{pmatrix} 3 & -4 & 1 \\ -4 & 0 & -4 \\ 1 & -4 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}
\]

(14.4)

which equals \( 3x^2 - 8xy + 2xz - 8yz + 3z^2 \). This is very awkward because of the mixed terms such as \(-8xy\). The idea is to pick different axes such that if \( x, y, z \) are taken with respect to these axes, the quadratic form is much simpler. In other words, look for new variables, \( x', y', \) and \( z' \) and a unitary matrix \( U \) such that

\[
U \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}
\]

(14.5)

and if you write the quadratic form in terms of the primed variables, there will be no mixed terms. Any symmetric real matrix is Hermitian and is therefore normal. From Corollary 14.1.6 it follows there exists a real unitary matrix \( U \) (an orthogonal matrix) such that \( U^T A U = D \) a diagonal matrix. Thus in the quadratic form, \( U^T A U = \)

\[
\begin{pmatrix} x' & y' & z' \end{pmatrix} \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix}
\]

(14.5)

and in terms of these new variables, the quadratic form becomes

\[
\lambda_1 (x')^2 + \lambda_2 (y')^2 + \lambda_3 (z')^2
\]

where \( D = \text{diag} (\lambda_1, \lambda_2, \lambda_3) \). Similar considerations apply equally well in any other dimension. For the given example,

\[
\begin{pmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} 3 & -4 & 1 \\ -4 & 0 & -4 \\ 1 & -4 & 3 \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix}
\]

\[
\begin{pmatrix} 2 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 8 \end{pmatrix}
\]
and so if the new variables are given by
\[
\begin{pmatrix}
\frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\
0 & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{3}} \\
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}}
\end{pmatrix}
\begin{pmatrix}
x' \\
y' \\
z'
\end{pmatrix} =
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix},
\]
it follows that in terms of the new variables the quadratic form is
\[2 \, (x')^2 - 4 \, (y')^2 + 8 \, (z')^2.\]
You can work other examples the same way.

### 14.3 The Estimation Of Eigenvalues

There are ways to estimate the eigenvalues for matrices. The most famous is known as Gerschgorin’s theorem. This theorem gives a rough idea where the eigenvalues are just from looking at the matrix.

**Theorem 14.3.1** Let \( A \) be an \( n \times n \) matrix. Consider the \( n \) Gerschgorin discs defined as
\[
D_i = \left\{ \lambda \in \mathbb{C} : |\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\}.
\]
Then every eigenvalue is contained in some Gerschgorin disc.

This theorem says to add up the absolute values of the entries of the \( i^{th} \) row which are off the main diagonal and form the disc centered at \( a_{ii} \) having this radius. The union of these discs contains \( \sigma (A) \).

**Proof:** Suppose \( A \mathbf{x} = \lambda \mathbf{x} \) where \( \mathbf{x} \neq 0 \). Then for \( A = (a_{ij}) \), let \( |x_k| \geq |x_j| \) for all \( x_j \). Thus \( |x_k| \neq 0 \).
\[
\sum_{j \neq k} a_{kj} x_j = (\lambda - a_{kk}) x_k.
\]
Then
\[
|x_k| \sum_{j \neq k} |a_{kj}| \geq \sum_{j \neq k} |a_{kj}| |x_j| \geq \sum_{j \neq k} a_{kj} x_j = |\lambda - a_{ii}| |x_k|.
\]
Now dividing by \( |x_k| \), it follows \( \lambda \) is contained in the \( k^{th} \) Gerschgorin disc. \( \blacksquare \)

**Example 14.3.2** Here is a matrix. Estimate its eigenvalues.
\[
\begin{pmatrix}
2 & 1 & 1 \\
3 & 5 & 0 \\
0 & 1 & 9
\end{pmatrix}
\]
According to Gerschgorin’s theorem the eigenvalues are contained in the disks
\[
D_1 = \{ \lambda \in \mathbb{C} : |\lambda - 2| \leq 2 \}, \quad D_2 = \{ \lambda \in \mathbb{C} : |\lambda - 5| \leq 3 \},
\]
\[D_3 = \{ \lambda \in \mathbb{C} : |\lambda - 9| \leq 1 \}
\]
It is important to observe that these disks are in the complex plane. In general this is the case. If you want to find eigenvalues they will be complex numbers.
So what are the values of the eigenvalues? In this case they are real. You can compute them by graphing the characteristic polynomial, \( \lambda^3 - 16 \lambda^2 + 70 \lambda - 66 \) and then zooming in on the zeros. If you do this you find the solution is \( \{ \lambda = 1.2953 \} , \{ \lambda = 5.5905 \} , \{ \lambda = 9.1142 \} \). Of course these are only approximations and so this information is useless for finding eigenvectors. However, in many applications, it is the size of the eigenvalues which is important and so these numerical values would be helpful for such applications. In this case, you might think there is no real reason for Gerschgorin’s theorem. Why not just compute the characteristic equation and graph and zoom? This is fine up to a point, but what if the matrix was huge? Then it might be hard to find the characteristic polynomial. Remember the difficulties in expanding a big matrix along a row or column. Also, what if the eigenvalues were complex? You don’t see these by following this procedure. However, Gerschgorin’s theorem will at least estimate them.

14.4 Advanced Theorems

More can be said but this requires some theory from complex variables. The following is a fundamental theorem about counting zeros.

**Theorem 14.4.1** Let \( U \) be a region and let \( \gamma : [a, b] \to U \) be closed, continuous, bounded variation, and the winding number, \( n(\gamma, z) = 0 \) for all \( z \notin U \). Suppose also that \( f \) is analytic on \( U \) having zeros \( a_1, \ldots, a_m \), where the zeros are repeated according to multiplicity, and suppose that none of these zeros are on \( \gamma([a, b]) \). Then

\[
\frac{1}{2\pi i} \int_\gamma \frac{f'(z)}{f(z)} \, dz = \sum_{k=1}^m n(\gamma, a_k).
\]

**Proof:** It is given that \( f(z) = \prod_{j=1}^m (z - a_j) g(z) \) where \( g(z) \neq 0 \) on \( U \). Hence using the product rule,

\[
\frac{f'(z)}{f(z)} = \sum_{j=1}^m \frac{1}{z - a_j} + \frac{g'(z)}{g(z)}
\]

where \( \frac{g'(z)}{g(z)} \) is analytic on \( U \) and so

\[
\frac{1}{2\pi i} \int_\gamma \frac{f'(z)}{f(z)} \, dz = \sum_{j=1}^m n(\gamma, a_j) + \frac{1}{2\pi i} \int_\gamma \frac{g'(z)}{g(z)} \, dz = \sum_{j=1}^m n(\gamma, a_j).
\]

Now let \( A \) be an \( n \times n \) matrix. Recall that the eigenvalues of \( A \) are given by the zeros of the polynomial, \( p_A(z) = \det(zI - A) \) where \( I \) is the \( n \times n \) identity. You can argue that small changes in \( A \) will produce small changes in \( p_A(z) \) and \( p_A'(z) \). Let \( \gamma_k \) denote a very small closed circle which winds around \( z_k \), one of the eigenvalues of \( A \), in the counter clockwise direction so that \( n(\gamma_k, z_k) = 1 \). This circle is to enclose only \( z_k \) and is to have no other eigenvalue on it. Then apply Theorem 14.4.1. According to this theorem

\[
\frac{1}{2\pi i} \int_\gamma \frac{p_A'(z)}{p_A(z)} \, dz
\]

is always an integer equal to the multiplicity of \( z_k \) as a root of \( p_A(t) \). Therefore, small changes in \( A \) result in no change to the above contour integral because it must be an integer and small changes in \( A \) result in small changes in the integral. Therefore whenever \( B \) is close enough to \( A \), the two matrices have the same number of zeros inside \( \gamma_k \), the zeros being counted according to multiplicity. By making the radius of the small circle equal to \( \varepsilon \) where \( \varepsilon \) is less than the minimum distance between any two distinct eigenvalues of \( A \), this shows that if \( B \) is close enough to \( A \), every eigenvalue of \( B \) is closer than \( \varepsilon \) to some eigenvalue of \( A \). □

\(^1\)If you haven’t studied the theory of a complex variable, you should skip this section because you won’t understand any of it.
Theorem 14.4.2 If \( \lambda \) is an eigenvalue of \( A \), then if all the entries of \( B \) are close enough to the corresponding entries of \( A \), some eigenvalue of \( B \) will be within \( \varepsilon \) of \( \lambda \).

Consider the situation that \( A(t) \) is an \( n \times n \) matrix and that \( t \to A(t) \) is continuous for \( t \in [0,1] \).

Lemma 14.4.3 Let \( \lambda(t) \in \sigma(A(t)) \) for \( t < 1 \) and let \( \Sigma_t = \cup_{s \geq t} \sigma(A(s)) \). Also let \( K_t \) be the connected component of \( \lambda(t) \) in \( \Sigma_t \). Then there exists \( \eta > 0 \) such that \( K_t \cap \sigma(A(s)) \neq \emptyset \) for all \( s \in [t,t+\eta] \).

Proof: Denote by \( D(\lambda(t), \delta) \) the disc centered at \( \lambda(t) \) having radius \( \delta > 0 \), with other occurrences of this notation being defined similarly. Thus

\[ D(\lambda(t), \delta) = \{ z \in \mathbb{C} : |\lambda(t) - z| \leq \delta \}. \]

Suppose \( \delta > 0 \) is small enough that \( \lambda(t) \) is the only element of \( \sigma(A(t)) \) contained in \( D(\lambda(t), \delta) \) and that \( p_{A(t)} \) has no zeroes on the boundary of this disc. Then by continuity, and the above discussion and theorem, there exists \( \eta > 0, t + \eta < 1 \), such that for \( s \in [t,t+\eta] \), \( p_{A(s)} \) also has no zeroes on the boundary of this disc and \( A(s) \) has the same number of eigenvalues, counted according to multiplicity, in the disc as \( A(t) \). Thus \( \sigma(A(s)) \cap D(\lambda(t), \delta) \neq \emptyset \) for all \( s \in [t,t+\eta] \). Now let

\[ H = \bigcup_{s \in [t,t+\eta]} \sigma(A(s)) \cap D(\lambda(t), \delta). \]

It will be shown that \( H \) is connected. Suppose not. Then \( H = P \cup Q \) where \( P, Q \) are separated and \( \lambda(t) \in P \). Let \( s_0 = \inf \{ s : \lambda(s) \in Q \text{ for some } \lambda(s) \in \sigma(A(s)) \} \). There exists \( \lambda(s_0) \in \sigma(A(s_0)) \cap D(\lambda(t), \delta) \). If \( \lambda(s_0) \notin Q \), then from the above discussion there are \( \lambda(s) \in \sigma(A(s)) \) such for \( s > s_0 \) arbitrarily close to \( \lambda(s_0) \). Therefore, \( \lambda(s_0) \in Q \) which shows that \( s_0 > t \) because \( \lambda(t) \) is the only element of \( \sigma(A(t)) \) in \( D(\lambda(t), \delta) \) and \( \lambda(t) \in P \). Now let \( s_n \uparrow s_0 \). Then \( \lambda(s_n) \in P \) for any \( \lambda(s_n) \in \sigma(A(s_n)) \cap D(\lambda(t), \delta) \) and also it follows from the above discussion that for some choice of \( s_n \to s_0, \lambda(s_n) \to \lambda(s_0) \) which contradicts \( P \) and \( Q \) separated and nonempty. Since \( P \) is nonempty, this shows \( Q = \emptyset \). Therefore, \( H \) is connected as claimed. But \( K_t \supset H \) and so \( K_t \cap \sigma(A(s)) \neq \emptyset \) for all \( s \in [t,t+\eta] \). ■

Theorem 14.4.4 Suppose \( A(t) \) is an \( n \times n \) matrix and that \( t \to A(t) \) is continuous for \( t \in [0,1] \). Let \( \lambda(0) \in \sigma(A(0)) \) and define \( \Sigma \equiv \cup_{t \in [0,1]} \sigma(A(t)) \). Let \( K_{\lambda(0)} = K_0 \) denote the connected component of \( \lambda(0) \) in \( \Sigma \). Then \( K_0 \cap \sigma(A(t)) \neq \emptyset \) for all \( t \in [0,1] \).

Proof: Let \( S \equiv \{ t \in [0,1] : K_0 \cap \sigma(A(s)) \neq \emptyset \text{ for all } s \in [0,t] \} \). Then \( 0 \in S \). Let \( t_0 = \sup(S) \). Say \( \sigma(A(t_0)) = \lambda_1(t_0), \ldots, \lambda_r(t_0) \).

Claim: At least one of these is a limit point of \( K_0 \) and consequently must be in \( K_0 \) which shows that \( S \) has a last point. Why is this claim true? Let \( s_n \uparrow t_0 \) so \( s_n \in S \). Now let the discs, \( D(\lambda_i(t_0), \delta), i = 1, \ldots, r \) be disjoint with \( p_{A(t_0)} \) having no zeroes on \( \gamma_i \) the boundary of \( D(\lambda_i(t_0), \delta) \). Then for \( n \) large enough it follows from Theorem 14.4.3 and the discussion following it that \( \sigma(A(s_n)) \) is contained in \( \cup_{i=1}^r D(\lambda_i(t_0), \delta) \). It follows that \( K_0 \cap (\sigma(A(t_0)) + D(0, \delta)) \neq \emptyset \) for all \( \delta \) small enough. This requires at least one of the \( \lambda_i(t_0) \) to be in \( K_0 \). Therefore, \( t_0 \in S \) and \( S \) has a last point.

Now by Lemma 14.4.3, if \( t_0 < 1 \), then \( K_0 \cup K_t \) would be a strictly larger connected set containing \( \lambda(0) \). (The reason this would be strictly larger is that \( K_0 \cap \sigma(A(s)) = \emptyset \) for some \( s \in (t,t+\eta) \) while \( K_t \cap \sigma(A(s)) \neq \emptyset \) for all \( s \in [t,t+\eta] \).) Therefore, \( t_0 = 1 \). ■

Corollary 14.4.5 Suppose one of the Gerschgorin discs, \( D_i \) is disjoint from the union of the others. Then \( D_i \) contains an eigenvalue of \( A \). Also, if there are \( n \) disjoint Gerschgorin discs, then each one contains an eigenvalue of \( A \).

Proof: Denote by \( A(t) \) the matrix \( (a_{ij}^t) \) where if \( i \neq j \), \( a_{ij}^t = ta_{ij} \) and \( a_{ii}^t = a_{ii} \). Thus to get \( A(t) \) multiply all non diagonal terms by \( t \). Let \( t \in [0,1] \). Then \( A(0) = \text{diag}(a_{11}, \ldots, a_{nn}) \) and \( A(1) = A \).
Furthermore, the map, \( t \rightarrow A(t) \) is continuous. Denote by \( D_j^t \) the Gerschgorin disc obtained from the \( j^{th} \) row for the matrix \( A(t) \). Then it is clear that \( D_j^t \subseteq D_j \) the \( j^{th} \) Gerschgorin disc for \( A \). It follows \( a_{ii} \) is the eigenvalue for \( A(0) \) which is contained in the disc, consisting of the single point \( a_{ii} \) which is contained in \( D_i \). Letting \( K \) be the connected component in \( \Sigma \) for \( \Sigma \) defined in Theorem 14.4.4 which is determined by \( a_{ii} \), Gerschgorin’s theorem implies that \( K \cap \sigma(A(t)) \subseteq \bigcup_{j=1}^{n} D_j^t \subseteq \bigcup_{j=1}^{n} D_j = D_i \cup (\bigcup_{j \neq i} D_j^t) \) and also, since \( K \) is connected, there are not points of \( K \) in both \( D_i \) and \((\bigcup_{j \neq i} D_j^t) \). Since at least one point of \( K \) is in \( D_i,(t_{ii}) \), it follows all of \( K \) must be contained in \( D_i \). Now by Theorem 14.4.5 this shows there are points of \( K \cap \sigma(A) \) in \( D_i \). The last assertion follows immediately.

This can be improved even more. This involves the following lemma.

**Lemma 14.4.6** In the situation of Theorem 14.4.5 suppose \( \lambda(0) = K_0 \cap \sigma(A(0)) \) and that \( \lambda(0) \) is a simple root of the characteristic equation of \( A(0) \). Then for all \( t \in [0,1] \),

\[
\sigma(A(t)) \cap K_0 = \lambda(t)
\]

where \( \lambda(t) \) is a simple root of the characteristic equation of \( A(t) \).

**Proof:** Let \( S = \{ t \in [0,1] : K_0 \cap \sigma(A(s)) = \lambda(s) \} \), a simple eigenvalue for all \( s \in [0,t] \). Then \( 0 \in S \) so it is nonempty. Let \( t_0 = \sup(S) \) and suppose \( \lambda_1 \neq \lambda_2 \) are two elements of \( \sigma(A(t_0)) \cap K_0 \). Then choosing \( \eta > 0 \) small enough, and letting \( D_i \) be disjoint discs containing \( \lambda_i \) respectively, similar arguments to those of Lemma 14.4.5 can be used to conclude

\[
H_i \equiv \bigcup_{s \in [t_0 - \eta, t_0]} \sigma(A(s)) \cap D_i
\]

is a connected and nonempty set for \( i = 1,2 \) which would require that \( H_i \subseteq K_0 \). But then there would be two different eigenvalues of \( A(s) \) contained in \( K_0 \), contrary to the definition of \( t_0 \). Therefore, there is at most one eigenvalue \( \lambda(t_0) \in K_0 \cap \sigma(A(t_0)) \). Could it be a repeated root of the characteristic equation? Suppose \( \lambda(t_0) \) is a repeated root of the characteristic equation. As before, choose a small disc, \( D \) centered at \( \lambda(t_0) \) and \( \eta \) small enough that

\[
H \equiv \bigcup_{s \in [t_0 - \eta, t_0]} \sigma(A(s)) \cap D
\]

is a nonempty connected set containing either multiple eigenvalues of \( A(s) \) or else a single repeated root to the characteristic equation of \( A(s) \). But since \( H \) is connected and contains \( \lambda(t_0) \) it must be contained in \( K_0 \) which contradicts the condition for \( s \in S \) for all these \( s \in [t_0 - \eta, t_0] \). Therefore, \( t_0 \in S \) as hoped. If \( t_0 < 1 \), there exists a small disc centered at \( \lambda(t_0) \) and \( \eta > 0 \) such that for all \( s \in [t_0, t_0 + \eta] \), \( A(s) \) has only simple eigenvalues in \( D \) and the only eigenvalues of \( A(s) \) which could be in \( K_0 \) are in \( D \). (This last assertion follows from noting that \( \lambda(t_0) \) is the only eigenvalue of \( A(t_0) \) in \( K_0 \) and so the others are at a positive distance from \( K_0 \). For \( s \) close enough to \( t_0 \), the eigenvalues of \( A(s) \) are either close to these eigenvalues of \( A(t_0) \) at a positive distance from \( K_0 \) or they are close to the eigenvalue \( \lambda(t_0) \) in which case it can be assumed they are in \( D \).) But this shows that \( t_0 \) is not really an upper bound to \( S \). Therefore, \( t_0 = 1 \) and the lemma is proved.

With this lemma, the conclusion of the above corollary can be sharpened.

**Corollary 14.4.7** Suppose one of the Gerschgorin discs, \( D_i \) is disjoint from the union of the others. Then \( D_i \) contains exactly one eigenvalue of \( A \) and this eigenvalue is a simple root to the characteristic polynomial of \( A \).

**Proof:** In the proof of Corollary 14.4.5 note that \( a_{ii} \) is a simple root of \( A(0) \) since otherwise the \( i^{th} \) Gerschgorin disc would not be disjoint from the others. Also, \( K \), the connected component determined by \( a_{ii} \) must be contained in \( D_i \) because it is connected and by Gerschgorin’s theorem above, \( K \cap \sigma(A(t)) \) must be contained in the union of the Gerschgorin discs. Since all the other eigenvalues of \( A(0) \), the \( a_{jj} \), are outside \( D_i \), it follows that \( K \cap \sigma(A(0)) = a_{ii} \). Therefore, by Lemma 14.4.5 \( K \cap \sigma(A(1)) = K \cap \sigma(A) \) consists of a single simple eigenvalue.
Example 14.4.8  Consider the matrix
\[
\begin{pmatrix}
  5 & 1 & 0 \\
  1 & 1 & 1 \\
  0 & 1 & 0
\end{pmatrix}
\]

The Gerschgorin discs are \( D(5,1), D(1,2), \) and \( D(0,1). \) Observe \( D(5,1) \) is disjoint from the other discs. Therefore, there should be an eigenvalue in \( D(5,1). \) The actual eigenvalues are not easy to find. They are the roots of the characteristic equation, \( t^3 - 6t^2 + 3t + 5 = 0. \) The numerical values of these are \(-0.669\ 66, 1.423\ 1, \) and \( 5.246\ 55, \) verifying the predictions of Gerschgorin’s theorem.

14.5 Exercises

1. Explain why it is typically impossible to compute the upper triangular matrix whose existence is guaranteed by Schur’s theorem.

2. Now recall the QR factorization of Theorem [13.2.3] on Page 298. The QR algorithm is a technique which does compute the upper triangular matrix in Schur’s theorem. There is much more to the QR algorithm than will be presented here. In fact, what I am about to show you is not the way it is done in practice. One first obtains what is called a Hessenburg matrix for which the algorithm will work better. However, the idea is as follows. Start with \( A \) an \( n \times n \) matrix having real eigenvalues. Form \( A = QR \) where \( Q \) is orthogonal and \( R \) is upper triangular. (Right triangular.) This can be done using the technique of Theorem [13.2.3] using Householder matrices. Next take \( A_1 = RQ. \) Show that \( A = QA_1Q^T. \) In other words these two matrices, \( A, A_1 \) are similar. Explain why they have the same eigenvalues. Continue by letting \( A_1 \) play the role of \( A. \) Thus the algorithm is of the form \( A_n = QR_n \) and \( A_{n+1} = R_{n+1}Q. \) Explain why \( A = Q_nA_nQ_n^T \) for some \( Q_n \) orthogonal. Thus \( A_n \) is a sequence of matrices each similar to \( A. \) The remarkable thing is that often these matrices converge to an upper triangular matrix \( T \) and \( A = QTQ^T \) for some orthogonal matrix, the limit of the \( Q_n \) where the limit means the entries converge. Then the process computes the upper triangular Schur form of the matrix \( A. \) Thus the eigenvalues of \( A \) appear on the diagonal of \( T. \) You will see approximately what these are as the process continues.

3. ↑Try the QR algorithm on
\[
\begin{pmatrix}
  -1 & -2 \\
   6 &  6
\end{pmatrix}
\]
which has eigenvalues 3 and 2. I suggest you use a computer algebra system to do the computations.

4. ↑Now try the QR algorithm on
\[
\begin{pmatrix}
  0 & -1 \\
   2 &  0
\end{pmatrix}
\]
Show that the algorithm cannot converge for this example. Hint: Try a few iterations of the algorithm. Use a computer algebra system if you like.

5. ↑Show the two matrices \( A \equiv \begin{pmatrix} 0 & -1 \\ 4 & 0 \end{pmatrix} \) and \( B \equiv \begin{pmatrix} 0 & -2 \\ 2 & 0 \end{pmatrix} \) are similar; that is there exists a matrix \( S \) such that \( A = S^{-1}BS \) but there is no orthogonal matrix \( Q \) such that \( Q^TBQ = A. \) Show the QR algorithm does converge for the matrix \( B \) although it fails to do so for \( A. \)

6. Let \( F \) be an \( m \times n \) matrix. Show that \( F^*F \) has all real eigenvalues and furthermore, they are all nonnegative.
7. If $A$ is a real $n \times n$ matrix and $\lambda$ is a complex eigenvalue $\lambda = a + ib, b \neq 0$, of $A$ having eigenvector $z + iw$, show that $w \neq 0$.

8. Suppose $A = Q^T D Q$ where $Q$ is an orthogonal matrix and all the matrices are real. Also $D$ is a diagonal matrix. Show that $A$ must be symmetric.

9. Suppose $A$ is an $n \times n$ matrix and there exists a unitary matrix $U$ such that

$$A = U^* DU$$

where $D$ is a diagonal matrix. Explain why $A$ must be normal.

10. If $A$ is Hermitian, show that $\det (A)$ must be real.

11. Show that every unitary matrix preserves distance. That is, if $U$ is unitary,

$$|Ux| = |x|.$$

12. Show that if a matrix does preserve distances, then it must be unitary.

13. Show that a complex normal matrix $A$ is unitary if and only if its eigenvalues have magnitude equal to 1.

14. Suppose $A$ is an $n \times n$ matrix which is diagonally dominant. Recall this means

$$\sum_{j \neq i} |a_{ij}| < |a_{ii}|$$

show $A^{-1}$ must exist.

15. Give some disks in the complex plane whose union contains all the eigenvalues of the matrix

$$\begin{pmatrix}
1 + 2i & 4 & 2 \\
0 & i & 3 \\
5 & 6 & 7
\end{pmatrix}$$

16. Show a square matrix is invertible if and only if it has no zero eigenvalues.

17. Using Schur’s theorem, show the trace of an $n \times n$ matrix equals the sum of the eigenvalues and the determinant of an $n \times n$ matrix is the product of the eigenvalues.

18. Using Schur’s theorem, show that if $A$ is any complex $n \times n$ matrix having eigenvalues $\{\lambda_i\}$ listed according to multiplicity, then $\sum_{i,j} |A_{ij}|^2 \geq \sum_{i=1}^n |\lambda_i|^2$. Show that equality holds if and only if $A$ is normal. This inequality is called Schur’s inequality. \[\square\]

19. Here is a matrix.

$$\begin{pmatrix}
1234 & 6 & 5 & 3 \\
0 & -654 & 9 & 123 \\
98 & 123 & 10,000 & 11 \\
56 & 78 & 98 & 400
\end{pmatrix}$$

I know this matrix has an inverse before doing any computations. How do I know?

20. Show the critical points of the following function are

$$(0, -3, 0), (2, -3, 0), \text{ and } \left(1, -3, -\frac{1}{3}\right)$$

and classify them as local minima, local maxima or saddle points.

$$f(x, y, z) = -\frac{3}{2}x^4 + 6x^3 - 6x^2 + zx^2 - 2zx - 2y^2 - 12y - 18 - \frac{3}{2}z^2.$$
21. Here is a function of three variables.

\[ f(x, y, z) = 13x^2 + 2xy + 8xz + 13y^2 + 8yz + 10z^2 \]

change the variables so that in the new variables there are no mixed terms, terms involving \( xy, yz \) etc. Two eigenvalues are 12 and 18.

22. Here is a function of three variables.

\[ f(x, y, z) = 2x^2 - 4x + 2 + 9yx - 9y - 3zx + 3z + 5y^2 - 9zy - 7z^2 \]

change the variables so that in the new variables there are no mixed terms, terms involving \( xy, yz \) etc. The eigenvalues of the matrix which you will work with are \(-\frac{17}{2}, \frac{19}{2}, -1\).

23. Here is a function of three variables.

\[ f(x, y, z) = -x^2 + 2xy + 2xz - y^2 + 2yz - z^2 + x \]

change the variables so that in the new variables there are no mixed terms, terms involving \( xy, yz \) etc.

24. Show the critical points of the function,

\[ f(x, y, z) = -2yx^2 - 6yx - 4zx^2 - 12zx + y^2 + 2yz. \]

are points of the form,

\[ (x, y, z) = (t, 2t^2 + 6t, -t^2 - 3t) \]

for \( t \in \mathbb{R} \) and classify them as local minima, local maxima or saddle points.

25. Show the critical points of the function

\[ f(x, y, z) = \frac{1}{2} x^4 - 4x^3 + 8x^2 - 3zx^2 + 12zx + 2y^2 + 4y + 2 + \frac{1}{2} z^2. \]

are \((0, -1, 0), (4, -1, 0),\) and \((2, -1, -12)\) and classify them as local minima, local maxima or saddle points.

26. Let \( f(x, y) = 3x^4 - 24x^2 + 48 - yx^2 + 4y. \) Find and classify the critical points using the second derivative test.

27. Let \( f(x, y) = 3x^4 - 5x^2 + 2 - y^2x^2 + y^2. \) Find and classify the critical points using the second derivative test.

28. Let \( f(x, y) = 5x^4 - 7x^2 - 2 - 3y^2x^2 + 11y^2 - 4y^4. \) Find and classify the critical points using the second derivative test.

29. Let \( f(x, y, z) = -2x^4 - 3yx^2 + 3x^2 + 5x^2z + 3y^2 - 6y + 3 - 3zy + 3z + z^2. \) Find and classify the critical points using the second derivative test.

30. Let \( f(x, y, z) = 3yx^2 - 3x^2 - x^2z - y^2 + 2y - 1 + 3zy - 3z - 3z^2. \) Find and classify the critical points using the second derivative test.

31. Let \( Q \) be orthogonal. Find the possible values of \( \det(Q) \).

32. Let \( U \) be unitary. Find the possible values of \( \det(U) \).

33. If a matrix is nonzero can it have only zero for eigenvalues?

34. A matrix \( A \) is called nilpotent if \( A^k = 0 \) for some positive integer \( k \). Suppose \( A \) is a nilpotent matrix. Show it has only 0 for an eigenvalue.
35. If \( A \) is a nonzero nilpotent matrix, show it must be defective.

36. Suppose \( A \) is a nondefective \( n \times n \) matrix and its eigenvalues are all either 0 or 1. Show \( A^2 = A \). Could you say anything interesting if the eigenvalues were all either 0, 1, or \(-1\)? By DeMoivre’s theorem, an \( n^{th} \) root of unity is of the form

\[
\left( \cos \left( \frac{2k\pi}{n} \right) + i \sin \left( \frac{2k\pi}{n} \right) \right)
\]

Could you generalize the sort of thing just described to get \( A^n = A \)? **Hint:** Since \( A \) is nondefective, there exists \( S \) such that \( S^{-1}AS = D \) where \( D \) is a diagonal matrix.

37. This and the following problems will present most of a differential equations course. Most of the explanations are given. You fill in any details needed. To begin with, consider the scalar initial value problem

\[
y' = ay, \quad y(t_0) = y_0
\]

When \( a \) is real, show the unique solution to this problem is \( y = y_0 e^{a(t-t_0)} \). Next suppose

\[
y' = (a + ib) y, \quad y(t_0) = y_0
\]  
(14.6)

where \( y(t) = u(t) + iv(t) \). Show there exists a unique solution and it is given by \( y(t) = y_0 e^{a(t-t_0)} (\cos b (t - t_0) + i \sin b (t - t_0)) \equiv e^{(a+ib)(t-t_0)} y_0 \).  
(14.7)

Next show that for a real or complex there exists a unique solution to the initial value problem

\[
y' = ay + f, \quad y(t_0) = y_0
\]

and it is given by

\[
y(t) = e^{a(t-t_0)} y_0 + e^{at} \int_{t_0}^t e^{-as} f(s) ds.
\]

**Hint:** For the first part write as \( y' - ay = 0 \) and multiply both sides by \( e^{-at} \). Then explain why you get

\[
\frac{d}{dt} (e^{-at} y(t)) = 0, \quad y(t_0) = 0.
\]

Now you finish the argument. To show uniqueness in the second part, suppose

\[
y' = (a + ib) y, \quad y(t_0) = 0
\]

and verify this requires \( y(t) = 0 \). To do this, note

\[
\overline{y}' = (a - ib) \overline{y}, \quad \overline{y}(t_0) = 0
\]

and that \( |y|^2(t_0) = 0 \) and

\[
\frac{d}{dt} |y(t)|^2 = y'(t) \overline{y}(t) + \overline{y}'(t) y(t)
\]

\[
= (a + ib) y(t) \overline{y}(t) + (a - ib) \overline{y}(t) y(t) = 2a |y(t)|^2.
\]

Thus from the first part \( |y(t)|^2 = 0 e^{-2at} = 0 \). Finally observe by a simple computation that 14.6 is solved by 14.7. For the last part, write the equation as

\[
y' - ay = f
\]

and multiply both sides by \( e^{-at} \) and then integrate from \( t_0 \) to \( t \) using the initial condition.
38. Now consider $A$ an $n \times n$ matrix. By Schur’s theorem there exists unitary $Q$ such that

$$Q^{-1}AQ = T$$

where $T$ is upper triangular. Now consider the first order initial value problem

$$x' = Ax, \quad x(t_0) = x_0.$$  

Show there exists a unique solution to this first order system. \textbf{Hint:} Let $y = Q^{-1}x$ and so the system becomes

$$y' = Ty, \quad y(t_0) = Q^{-1}x_0$$  \hspace{1cm} (14.8)

Now letting $y = (y_1, \cdots, y_n)^T$, the bottom equation becomes

$$y'_n = t_{nn}y_n, \quad y_n(t_0) = (Q^{-1}x_0)_n.$$  

Then use the solution you get in this to get the solution to the initial value problem which occurs one level up, namely

$$y'_{n-1} = t_{(n-1)(n-1)}y_{n-1} + t_{(n-1)n}y_n, \quad y_{n-1}(t_0) = (Q^{-1}x_0)_{n-1}$$

Continue doing this to obtain a unique solution to \textbf{ES}.  

39. Now suppose $\Phi(t)$ is an $n \times n$ matrix of the form

$$\Phi(t) = \left( \begin{array}{cccc} x_1(t) & \cdots & x_n(t) \end{array} \right)$$  

where

$$x'_k(t) = Ax_k(t).$$  \hspace{1cm} (14.9)

Explain why

$$\Phi'(t) = A\Phi(t)$$

if and only if $\Phi(t)$ is given in the form of \textbf{ES}. Also explain why if $c \in \mathbb{F}^n$, $y(t) = \Phi(t)c$ solves the equation $y'(t) = Ay(t)$.  

40. In the above problem, consider the question whether all solutions to

$$x' = Ax$$  \hspace{1cm} (14.10)

are obtained in the form $\Phi(t)c$ for some choice of $c \in \mathbb{F}^n$. In other words, is the general solution to this equation $\Phi(t)c$ for $c \in \mathbb{F}^n$? Prove the following theorem using linear algebra.  

\textbf{Theorem 14.5.1} Suppose $\Phi(t)$ is an $n \times n$ matrix which satisfies $\Phi'(t) = A\Phi(t)$. Then the general solution to \textbf{ES} is $\Phi(t)c$ if and only if $\Phi(t)^{-1}$ exists for some $t$. Furthermore, if $\Phi'(t) = A\Phi(t)$, then either $\Phi(t)^{-1}$ exists for all $t$ or $\Phi(t)^{-1}$ never exists for any $t$.  

($\det(\Phi(t))$ is called the Wronskian and this theorem is sometimes called the Wronskian alternative.)  

\textbf{Hint:} Suppose first the general solution is of the form $\Phi(t)c$ where $c$ is an arbitrary constant vector in $\mathbb{F}^n$. You need to verify $\Phi(t)^{-1}$ exists for some $t$. In fact, show $\Phi(t)^{-1}$ exists for every $t$. Suppose then that $\Phi(t_0)^{-1}$ does not exist. Explain why there exists $c \in \mathbb{F}^n$ such that there is no solution $x$ to the equation $c = \Phi(t_0)x$. By the existence part of Problem \textbf{ES} there exists a solution to

$$x' = Ax, \quad x(t_0) = c$$

but this cannot be in the form $\Phi(t)c$. Thus for every $t$, $\Phi(t)^{-1}$ exists. Next suppose for some $t_0$, $\Phi(t_0)^{-1}$ exists. Let $z' = Az$ and choose $c$ such that

$$z(t_0) = \Phi(t_0)c$$
Then both \( z(t), \phi(t)c \) solve
\[
x' = Ax, \quad x(t_0) = z(t_0)
\]
Apply uniqueness to conclude \( z = \phi(t)c \). Finally, consider that \( \phi(t)c \) for \( c \in \mathbb{F}^n \) either is the general solution or it is not the general solution. If it is, then \( \phi(t)^{-1} \) exists for all \( t \). If it is not, then \( \phi(t)^{-1} \) cannot exist for any \( t \) from what was just shown.

41. Let \( \phi'(t) = A\phi(t) \). Then \( \phi(t) \) is called a fundamental matrix if \( \phi(t)c \) for \( c \in \mathbb{F}^n \) either is the general solution or it is not the general solution. If it is, then \( \phi(t)c \) exists for all \( t \); if it is not, then \( \phi(t)c \) cannot exist for any \( t \) from what was just shown.

42. Show there exists a special \( \phi \) such that \( \phi'(t) = A\phi(t) \), \( \phi(0) = I \), and suppose \( \phi(t)c \) exists for all \( t \). Show using uniqueness that
\[
\phi(-t) = \phi(t)^{-1}
\]
and that for all \( t, s \in \mathbb{R} \)
\[
\phi(t+s) = \phi(t)\phi(s)
\]
Explain why with this special \( \phi \), the solution to (14.11) can be written as
\[
x(t) = \phi(t - t_0)x_0 + \int_{t_0}^{t} \phi(t - s)f(s)ds.
\]

43. You can see more on this problem and the next one in the latest version of Horn and Johnson, [13]. Two \( m \times n \) matrices \( A, B \) are said to be congruent if there is an invertible \( P \) such that
\[
B = PAP^*
\]
Let \( A \) be a Hermitian matrix. Thus it has all real eigenvalues. Let \( n_+ \) be the number of positive eigenvalues, \( n_- \), the number of negative eigenvalues and \( n_0 \) the number of zero eigenvalues. For \( k \) a positive integer, let \( I_k \) denote the \( k \times k \) identity matrix and \( O_k \) the \( k \times k \) zero matrix. Then the inertia matrix of \( A \) is the following block diagonal \( m \times n \) matrix.
\[
\begin{pmatrix}
I_{n_+} & I_{n_-} \\
I_{n_-} & O_{n_0}
\end{pmatrix}
\]
Show that $A$ is congruent to its inertia matrix. Next show that congruence is an equivalence relation on the set of Hermitian matrices. Finally, show that if two Hermitian matrices have the same inertia matrix, then they must be congruent. **Hint:** First recall that there is a unitary matrix, $U$ such that

$$U^*A U = \begin{pmatrix} D_{n_+} & 0 \\ 0 & D_{n_-} \end{pmatrix}$$

where the $D_{n_+}$ is a diagonal matrix having the positive eigenvalues of $A$, $D_{n_-}$ being defined similarly. Now let $|D_{n_-}|$ denote the diagonal matrix which replaces each entry of $D_{n_-}$ with its absolute value. Consider the two diagonal matrices

$$D = D^* = \begin{pmatrix} D_{n_+}^{-1/2} & |D_{n_-}|^{-1/2} \\ |D_{n_-}|^{-1/2} & I_{n_0} \end{pmatrix}$$

Now consider $D^*U^*AUD$.

44. Show that if $A, B$ are two congruent Hermitian matrices, then they have the same inertia matrix. **Hint:** Let $A = SBS^*$ where $S$ is invertible. Show that $A, B$ have the same rank and this implies that they are each unitarily similar to a diagonal matrix which has the same number of zero entries on the main diagonal. Therefore, letting $V_A$ be the span of the eigenvectors associated with positive eigenvalues of $A$ and $V_B$ being defined similarly, it suffices to show that these have the same dimensions. Show that $(Ax, x) > 0$ for all $x \in V_A$. Next consider $S^*V_A$. For $x \in V_A$, explain why

$$(BS^*x, S^*x) = \left( S^{-1}A(S^*)^{-1} S^*x, S^*x \right)$$

$$= (S^{-1}Ax, S^*x) = \left( Ax, (S^{-1})^* S^*x \right) = (Ax, x) > 0$$

Next explain why this shows that $S^*V_A$ is a subspace of $V_B$ and so the dimension of $V_B$ is at least as large as the dimension of $V_A$. Hence there are at least as many positive eigenvalues for $B$ as there are for $A$. Switching $A, B$ you can turn the inequality around. Thus the two have the same inertia matrix.

45. Let $A$ be an $m \times n$ matrix. Then if you unraveled it, you could consider it as a vector in $\mathbb{C}^{nm}$. The Frobenius inner product on the vector space of $m \times n$ matrices is defined as

$$(A, B) \equiv \text{trace}(AB^*)$$

Show that this really does satisfy the axioms of an inner product space and that it also amounts to nothing more than considering $m \times n$ matrices as vectors in $\mathbb{C}^{nm}$.

46. \*Consider the $n \times n$ unitary matrices. Show that whenever $U$ is such a matrix, it follows that

$$|U|_{\mathbb{C}^{n \times n}} = \sqrt{n}$$

Next explain why if $\{U_k\}$ is any sequence of unitary matrices, there exists a subsequence $\{U_{km}\}_{m=1}^\infty$ such that $\lim_{m \to \infty} U_{km} = U$ where $U$ is unitary. Here the limit takes place in the sense that the entries of $U_{km}$ converge to the corresponding entries of $U$.

47. \*Let $A, B$ be two $n \times n$ matrices. Denote by $\sigma(A)$ the set of eigenvalues of $A$. Define

$$\text{dist}(\sigma(A), \sigma(B)) = \max_{\lambda \in \sigma(A)} \min \{|\lambda - \mu| : \mu \in \sigma(B)\}$$

Explain why $\text{dist}(\sigma(A), \sigma(B))$ is small if and only if every eigenvalue of $A$ is close to some eigenvalue of $B$. Now prove the following theorem using the above problem and Schur’s theorem. This theorem says roughly that if $A$ is close to $B$ then the eigenvalues of $A$ are close to those of $B$ in the sense that every eigenvalue of $A$ is close to an eigenvalue of $B$. 

Theorem 14.5.2 Suppose \( \lim_{k \to \infty} A_k = A \). Then
\[
\lim_{k \to \infty} \text{dist} (\sigma(A_k), \sigma(A)) = 0
\]

48. Let \( A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \) be a \( 2 \times 2 \) matrix which is not a multiple of the identity. Show that \( A \) is similar to a \( 2 \times 2 \) matrix which has at least one diagonal entry equal to 0. **Hint:** First note that there exists a vector \( \mathbf{a} \) such that \( A\mathbf{a} \) is not a multiple of \( \mathbf{a} \). Then consider
\[
B = \left( \begin{array}{cc} a & A\mathbf{a} \\ \mathbf{a} & A \end{array} \right)^{-1} A \left( \begin{array}{cc} a & A\mathbf{a} \\ \mathbf{a} & A \end{array} \right)
\]
Show \( B \) has a zero on the main diagonal.

49. Let \( A \) be a complex \( n \times n \) matrix which has trace equal to 0. Show that \( A \) is similar to a matrix which has all zeros on the main diagonal. **Hint:** Use Problem 34 on Page 143 to argue that you can say that a given matrix is similar to one which has the diagonal entries permuted in any order desired. Then use the above problem and block multiplication to show that if the \( A \) has \( k \) nonzero entries, then it is similar to a matrix which has \( k-1 \) nonzero entries. Finally, when \( A \) is similar to one which has at most one nonzero entry, this one must also be zero because of the condition on the trace.

50. An \( n \times n \) matrix \( X \) is a comutator if there are \( n \times n \) matrices \( A, B \) such that \( X = AB - BA \). Show that the trace of any comutator is 0. Next show that if a complex matrix \( X \) has trace equal to 0, then it is in fact a comutator. **Hint:** Use the above problem to show that it suffices to consider \( X \) having all zero entries on the main diagonal. Then define
\[
A = \begin{pmatrix} 1 & 0 & \cdots \\ 2 & \ddots & \vdots \\ 0 & \cdots & n \end{pmatrix}, \quad B_{ij} = \begin{cases} \frac{x_{ij}}{i-j} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}
\]

14.6 The Right Polar Factorization*

The right polar factorization involves writing a matrix as a product of two other matrices, one which preserves distances and the other which stretches and distorts. This is of fundamental significance in geometric measure theory and also in continuum mechanics. Not surprisingly the stress should depend on the part which stretches and distorts. See [13].

First here are some lemmas which review and add to many of the topics discussed so far about adjoints and orthonormal sets and such things.

**Lemma 14.6.1** Let \( A \) be a Hermitian matrix such that all its eigenvalues are nonnegative. Then there exists a Hermitian matrix \( A^{1/2} \) such that \( A^{1/2} \) has all nonnegative eigenvalues and \( (A^{1/2})^2 = A \).

**Proof:** Since \( A \) is Hermitian, there exists a diagonal matrix \( D \) having all real nonnegative entries and a unitary matrix \( U \) such that \( A = U^*DU \). Then denote by \( D^{1/2} \) the matrix which is obtained by replacing each diagonal entry of \( D \) with its square root. Thus \( D^{1/2}D^{1/2} = D \). Then define
\[
A^{1/2} = U^*D^{1/2}U.
\]
Then
\[
\left( A^{1/2} \right)^2 = U^*D^{1/2}UU^*D^{1/2}U = U^*DU = A.
\]
Since $D^{1/2}$ is real,
\[ \left( U^* D^{1/2} U \right)^* = U^* \left( D^{1/2} \right)^* (U^*)^* = U^* D^{1/2} U \]
so $A^{1/2}$ is Hermitian. \[\square\]

Next it is helpful to recall the Gram Schmidt algorithm and observe a certain property stated in the next lemma.

**Lemma 14.6.2** Suppose \{w₁, · · · , w₉, v₉₊₁, · · · , vₚ \} is a linearly independent set of vectors such that \{w₁, · · · , w₉ \} is an orthonormal set of vectors. Then when the Gram Schmidt process is applied to the vectors in the given order, it will not change any of the w₁, · · · , w₉.

**Proof:** Let \{u₁, · · · , uₚ \} be the orthonormal set delivered by the Gram Schmidt process. Then \( u₁ = w₁ \) because by definition, \( u₁ \equiv w₁/|w₁| = w₁ \). Now suppose \( u_j = w_j \) for all \( j \leq k \leq r \). Then if \( k < r \), consider the definition of \( u_{k+1} \).

\[ u_{k+1} = \frac{w_{k+1} - \sum_{j=1}^{k+1} (w_{k+1}, u_j) u_j}{\left\| w_{k+1} - \sum_{j=1}^{k+1} (w_{k+1}, u_j) u_j \right\|} \]

By induction, \( u_j = w_j \) and so this reduces to \( w_{k+1}/|w_{k+1}| = w_{k+1} \). \[\square\]

This lemma immediately implies the following lemma.

**Lemma 14.6.3** Let \( V \) be a subspace of dimension \( p \) and let \{w₁, · · · , w₉ \} be an orthonormal set of vectors in \( V \). Then this orthonormal set of vectors may be extended to an orthonormal basis for \( V \),

\[ \{w₁, · · · , w₉, v_{₉₊₁}, · · · , vₚ \} \]

**Proof:** First extend the given linearly independent set \{w₁, · · · , w₉ \} to a basis for \( V \) and then apply the Gram Schmidt theorem to the resulting basis. Since \{w₁, · · · , w₉ \} is orthonormal it follows from Lemma 14.6.2 the result is of the desired form, an orthonormal basis extending \{w₁, · · · , w₉ \}.

Recall Lemma 14.6.2 which is about preserving distances. It is restated here in the case of an \( m \times n \) matrix.

**Lemma 14.6.4** Suppose \( R \) is an \( m \times n \) matrix with \( m > n \) and \( R \) preserves distances. Then \( R^* R = I \).

With this preparation, here is the big theorem about the right polar factorization.

**Theorem 14.6.5** Let \( F \) be an \( m \times n \) matrix where \( m \geq n \). Then there exists a Hermitian \( n \times n \) matrix \( U \) which has all nonnegative eigenvalues and an \( m \times n \) matrix \( R \) which preserves distances and satisfies \( R^* R = I \) such that \( F = RU \).

**Proof:** Consider \( F^* F \). This is a Hermitian matrix because

\[ (F^* F)^* = F^* (F^*)^* = F^* F \]

Also the eigenvalues of the \( n \times n \) matrix \( F^* F \) are all nonnegative. This is because if \( x \) is an eigenvalue,

\[ \lambda(x, x) = (F^* F x, x) = (Fx, Fx) \geq 0. \]

Therefore, by Lemma 14.6.3, there exists an \( n \times n \) Hermitian matrix \( U \) having all nonnegative eigenvalues such that

\[ U^2 = F^* F. \]

Consider the subspace \( U (\mathbb{F}^n) \). Let \( \{Ux₁, · · · , Uxₙ \} \) be an orthonormal basis for

\[ U (\mathbb{F}^n) \subseteq \mathbb{F}^n. \]
Note that $U (\mathbb{F}^n)$ might not be all of $\mathbb{F}^n$. Using Lemma 14.6.3, extend to an orthonormal basis for all of $\mathbb{F}^n$,
\[
\{Ux_1, \ldots, Ux_r, y_{r+1}, \ldots, y_n\}.
\]
Next observe that $\{Fx_1, \ldots, Fx_r\}$ is also an orthonormal set of vectors in $\mathbb{F}^m$. This is because
\[
(Fx_k, Fx_j) = (F^*Fx_k, x_j) = (U^2x_k, x_j) = (Ux_k, U^*x_j) = \delta_{jk}
\]
Therefore, from Lemma 14.6.3, again, this orthonormal set of vectors can be extended to an orthonormal basis for $\mathbb{F}^m$,
\[
\{Fx_1, \ldots, Fx_r, z_{r+1}, \ldots, z_m\}
\]
Thus there are at least as many $z_k$ as there are $y_j$ because $m \geq n$. Now for $x \in \mathbb{F}^n$, since
\[
\{Ux_1, \ldots, Ux_r, y_{r+1}, \ldots, y_n\}
\]
is an orthonormal basis for $\mathbb{F}^n$, there exist unique scalars,
\[
c_1 \ldots, c_r, d_{r+1}, \ldots, d_n
\]
such that
\[
x = \sum_{k=1}^r c_k Ux_k + \sum_{k=r+1}^n d_k y_k
\]
Define
\[
Rx \equiv \sum_{k=1}^r c_k Fx_k + \sum_{k=r+1}^n d_k z_k \quad (14.12)
\]
Then also there exist scalars $b_k$ such that
\[
Ux = \sum_{k=1}^r b_k Ux_k
\]
and so from 14.6.3,
\[
RUx = \sum_{k=1}^r b_k Fx_k = F \left( \sum_{k=1}^r b_k x_k \right)
\]
Is $F (\sum_{k=1}^r b_k x_k) = F (x)$?
\[
\left( F \left( \sum_{k=1}^r b_k x_k \right) - F (x), F \left( \sum_{k=1}^r b_k x_k \right) - F (x) \right)
\]
\[
= \left( (F^*F) \left( \sum_{k=1}^r b_k x_k - x \right), \left( \sum_{k=1}^r b_k x_k - x \right) \right)
\]
\[
= \left( U^2 \left( \sum_{k=1}^r b_k x_k - x \right), \left( \sum_{k=1}^r b_k x_k - x \right) \right)
\]
\[
= \left( U \left( \sum_{k=1}^r b_k x_k - x \right), U \left( \sum_{k=1}^r b_k x_k - x \right) \right)
\]
\[
= \left( \sum_{k=1}^r b_k Ux_k - Ux, \sum_{k=1}^r b_k Ux_k - Ux \right) = 0
\]
Therefore, $F (\sum_{k=1}^r b_k x_k) = F (x)$ and this shows $RUx = Fx$. Thus
\[
|RUx|^2 = |Fx|^2 = (F^*Fx, x) = |Ux|^2
\]
and so $R$ preserves distances. Therefore, by Lemma 14.6.3, $R^*R = I$.

Then there is an easy corollary of this theorem.
Corollary 14.6.6 Let $F$ be $m \times n$ and suppose $n \geq m$. Then there exists a Hermitian $U$ and and $R$, such that 

$$F = UR, \quad RR^* = I.$$ 

Proof: Recall that $L^* = L$ and $(ML)^* = L^*M^*$. Now apply Theorem 14.6.5 to $F^*$. Thus, $F^* = R^*U$ where $R^*$ and $U$ satisfy the conditions of that theorem. In particular $R^*$ preserves distances. Then $F = UR$ and $RR^* = R^*R^* = I$. \[\blacksquare\]

14.7 An Application To Statistics

A random vector is a function $X : \Omega \to \mathbb{R}^p$ where $\Omega$ is a probability space. This means that there exists an algebra of measurable sets $\mathcal{F}$ and a probability measure $P : \mathcal{F} \to [0, 1]$. In practice, people often don’t worry too much about the underlying probability space and instead pay more attention to the distribution measure of the random variable. For $E$ a suitable subset of $\mathbb{R}^p$, this measure gives the probability that $X$ has values in $E$. There are often excellent reasons for believing that a random vector is normally distributed. This means that the probability that $X$ has values in a set $E$ is given by

$$\int_E \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)^* \Sigma^{-1} (x - \mu) \right) dx$$

The expression in the integral is called the normal probability density function. There are two parameters, $\mu$ and $\Sigma$ where $\mu$ is called the mean and $\Sigma$ is called the covariance matrix. It is a symmetric matrix which has all real eigenvalues which are all positive. While it may be reasonable to assume this is the distribution, in general, you won’t know $\mu$ and $\Sigma$ and in order to use this formula to predict anything, you would need to know these quantities. I am following a nice discussion given in Wikipedia which makes use of the existence of square roots.

What people do to estimate $\mu$, and $\Sigma$ is to take $n$ independent observations $x_1, \cdots, x_n$ and try to predict what $\mu$ and $\Sigma$ should be based on these observations. One criterion used for making this determination is the method of maximum likelihood. In this method, you seek to choose the two parameters in such a way as to maximize the likelihood which is given as

$$\prod_{i=1}^n \frac{1}{\det(\Sigma)^{1/2}} \exp \left( -\frac{1}{2} (x_i - \mu)^* \Sigma^{-1} (x_i - \mu) \right).$$

For convenience the term $(2\pi)^{p/2}$ was ignored. Maximizing the above is equivalent to maximizing the ln of the above. So taking ln,

$$\frac{n}{2} \ln(\det(\Sigma^{-1})) - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^* \Sigma^{-1} (x_i - \mu)$$

Note that the above is a function of the entries of $\mu$. Take the partial derivative with respect to $m_l$. Since the matrix $\Sigma^{-1}$ is symmetric this implies

$$\sum_{i=1}^n \sum_r (x_{ir} - m_r) \Sigma_{rl}^{-1} = 0 \text{ each } l.$$ 

Written in terms of vectors,

$$\sum_{i=1}^n (x_i - \mu)^* \Sigma^{-1} = 0$$

and so, multiplying by $\Sigma$ on the right and then taking adjoints, this yields

$$\sum_{i=1}^n (x_i - \mu) = 0, \quad nm = \sum_{i=1}^n x_i, \quad \mu = \frac{1}{n} \sum_{i=1}^n x_i \equiv \bar{x}.$$
Now that \( \mathbf{m} \) is determined, it remains to find the best estimate for \( \Sigma \). \( (\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1} (\mathbf{x}_i - \mathbf{m}) \) is a scalar, so since \( \text{trace} (\mathbf{A}\mathbf{B}) = \text{trace} (\mathbf{B}\mathbf{A}) \),

\[
(\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1} (\mathbf{x}_i - \mathbf{m}) = \text{trace} ((\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1} (\mathbf{x}_i - \mathbf{m})) = \text{trace} ((\mathbf{x}_i - \mathbf{m}) (\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1})
\]

Therefore, the thing to maximize is

\[
n \ln (\det (\Sigma^{-1})) - \sum_{i=1}^{n} \text{trace} ((\mathbf{x}_i - \mathbf{m}) (\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1})
\]

\[
= n \ln (\det (\Sigma^{-1})) - \text{trace} \left( \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}) (\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1} \right)
\]

We assume that \( \mathbf{S} \) has rank \( p \). Thus it is a self adjoint matrix which has all positive eigenvalues. Therefore, from the property of the trace, \( \text{trace} (\mathbf{A}\mathbf{B}) = \text{trace} (\mathbf{B}\mathbf{A}) \), the thing to maximize is

\[
n \ln (\det (\Sigma^{-1})) - \text{trace} \left( \mathbf{S}^{1/2} (\mathbf{S}^{-1})^{1/2} \right)
\]

Now let \( \mathbf{B} = \mathbf{S}^{1/2} \Sigma^{-1} \mathbf{S}^{1/2} \). Then \( \mathbf{B} \) is positive and self adjoint also and so there exists \( \mathbf{U} \) unitary such that \( \mathbf{B} = \mathbf{U}^* \mathbf{D} \mathbf{U} \) where \( \mathbf{D} \) is the diagonal matrix having the positive scalars \( \lambda_1, \ldots, \lambda_p \) down the main diagonal. Solving for \( \Sigma^{-1} \) in terms of \( \mathbf{B} \), this yields \( \mathbf{S}^{-1/2} \mathbf{B} \mathbf{S}^{-1/2} = \Sigma^{-1} \) and so

\[
\ln (\det (\Sigma^{-1})) = \ln (\det (\mathbf{S}^{-1/2}) \det (\mathbf{D}) \det (\mathbf{S}^{-1/2})) = \ln (\det (\mathbf{S}^{-1})) + \ln (\det (\mathbf{D}))
\]

which yields

\[
C(S) + n \ln (\det (\mathbf{D})) - \text{trace} (\mathbf{D})
\]

as the thing to maximize. Of course this yields

\[
C(S) + n \ln \left( \prod_{i=1}^{p} \lambda_i \right) - \sum_{i=1}^{p} \lambda_i
\]

\[
= C(S) + n \sum_{i=1}^{p} \ln (\lambda_i) - \sum_{i=1}^{p} \lambda_i
\]

as the quantity to be maximized. To do this, take \( \partial / \partial \lambda_k \) and set equal to 0. This yields \( \lambda_k = n \). Therefore, from the above, \( \mathbf{B} = \mathbf{U}^* \mathbf{n} \mathbf{I} \mathbf{U} = n \mathbf{I} \). Also from the above,

\[
\mathbf{B}^{-1} = \frac{-1}{n} \mathbf{I} = \mathbf{S}^{-1/2} \Sigma \mathbf{S}^{-1/2}
\]

and so

\[
\Sigma = \frac{-1}{n} \mathbf{S} = \frac{-1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}) (\mathbf{x}_i - \mathbf{m})^*
\]

This has shown that the maximum likelihood estimates are

\[
\mathbf{m} = \bar{\mathbf{x}} \equiv \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i, \quad \Sigma = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}) (\mathbf{x}_i - \mathbf{m})^*.
\]
14.8 Simultaneous Diagonalization

Recall the following definition of what it means for a matrix to be diagonalizable.

**Definition 14.8.1** Let $A$ be an $n \times n$ matrix. It is said to be diagonalizable if there exists an invertible matrix $S$ such that

$$S^{-1}AS = D$$

where $D$ is a diagonal matrix.

Also, here is a useful observation.

**Observation 14.8.2** If $A$ is an $n \times n$ matrix and $AS = SD$ for $D$ a diagonal matrix, then each column of $S$ is an eigenvector or else it is the zero vector. This follows from observing that for $s_k$ the $k^{th}$ column of $S$ and from the way we multiply matrices,

$$As_k = \lambda_k s_k$$

It is sometimes interesting to consider the problem of finding a single similarity transformation which will diagonalize all the matrices in some set.

**Lemma 14.8.3** Let $A$ be an $n \times n$ matrix and let $B$ be an $m \times m$ matrix. Denote by $C$ the matrix

$$C \equiv \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}.$$  

Then $C$ is diagonalizable if and only if both $A$ and $B$ are diagonalizable.

**Proof:** Suppose $S_A^{-1}AS_A = D_A$ and $S_B^{-1}BS_B = D_B$ where $D_A$ and $D_B$ are diagonal matrices. You should use block multiplication to verify that $S \equiv \begin{pmatrix} S_A & 0 \\ 0 & S_B \end{pmatrix}$ is such that $S^{-1}CS = D_C$, a diagonal matrix.

Conversely, suppose $C$ is diagonalized by $S = (s_1, \ldots, s_{n+m})$. Thus $S$ has columns $s_i$. For each of these columns, write in the form

$$s_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$$

where $x_i \in \mathbb{F}^n$ and where $y_i \in \mathbb{F}^m$. The result is

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

where $S_{11}$ is an $n \times n$ matrix and $S_{22}$ is an $m \times m$ matrix. Then there is a diagonal matrix, $D_1$ being $n \times n$ and $D_2$ $m \times m$ such that

$$D = \text{diag}(\lambda_1, \ldots, \lambda_{n+m}) = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}$$

such that

$$\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}$$
Hence by block multiplication

\[ AS_{11} = S_{11}D_1, \quad BS_{22} = S_{22}D_2 \]

\[ BS_{21} = S_{21}D_1, \quad AS_{12} = S_{12}D_2 \]

It follows each of the \( x_i \) is an eigenvector of \( A \) or else is the zero vector and that each of the \( y_i \) is an eigenvector of \( B \) or is the zero vector. If there are \( n \) linearly independent \( x_i \), then \( A \) is diagonalizable by Theorem 6.3.6 on Page 6.3.6.

The row rank of the matrix \((x_1, \cdots, x_{n+m})\) must be \( n \) because if this is not so, the rank of \( S \) would be less than \( n + m \) which would mean \( S^{-1} \) does not exist. Therefore, since the column rank equals the row rank, this matrix has column rank equal to \( n \) and this means there are \( n \) linearly independent eigenvectors of \( A \) implying that \( A \) is diagonalizable. Similar reasoning applies to \( B \).

The following corollary follows from the same type of argument as the above.

**Corollary 14.8.4** Let \( A_k \) be an \( n_k \times n_k \) matrix and let \( C \) denote the block diagonal

\[
\left( \sum_{k=1}^{r} n_k \right) \times \left( \sum_{k=1}^{r} n_k \right)
\]

matrix given below.

\[
C \equiv \begin{pmatrix}
A_1 & 0 \\
& \ddots \\
0 & A_r
\end{pmatrix}
\]

Then \( C \) is diagonalizable if and only if each \( A_k \) is diagonalizable.

**Definition 14.8.5** A set, \( F \) of \( n \times n \) matrices is said to be simultaneously diagonalizable if and only if there exists a single invertible matrix \( S \) such that for every \( A \in F \), \( S^{-1}AS = D_A \) where \( D_A \) is a diagonal matrix. \( F \) is a commuting family of matrices if whenever \( A, B \in F \), \( AB = BA \).

**Lemma 14.8.6** If \( F \) is a set of \( n \times n \) matrices which is simultaneously diagonalizable, then \( F \) is a commuting family of matrices.

**Proof:** Let \( A, B \in F \) and let \( S \) be a matrix which has the property that \( S^{-1}AS \) is a diagonal matrix for all \( A \in F \). Then \( S^{-1}AS = D_A \) and \( S^{-1}BS = D_B \) where \( D_A \) and \( D_B \) are diagonal matrices. Since diagonal matrices commute,

\[
AB = SD_A S^{-1}SD_B S^{-1} = SD_B D_A S^{-1} = SD_B S^{-1}SD_A S^{-1} = BA.
\]

**Lemma 14.8.7** Let \( D \) be a diagonal matrix of the form

\[
D \equiv \begin{pmatrix}
\lambda_1 I_{n_1} & 0 & \cdots & 0 \\
0 & \lambda_2 I_{n_2} & \cdots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & \lambda_r I_{n_r}
\end{pmatrix}, \quad (14.13)
\]

where \( I_{n_i} \) denotes the \( n_i \times n_i \) identity matrix and \( \lambda_i \neq \lambda_j \) for \( i \neq j \) and suppose \( B \) is a matrix which commutes with \( D \). Then \( B \) is a block diagonal matrix of the form

\[
B = \begin{pmatrix}
B_1 & 0 & \cdots & 0 \\
0 & B_2 & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & B_r
\end{pmatrix}, \quad (14.14)
\]

where \( B_i \) is an \( n_i \times n_i \) matrix.
**Proof:** Let $B = (B_{ij})$ where $B_{ii} = B_i$ a block matrix as above in 14.14.

\[
\begin{pmatrix}
B_{11} & B_{12} & \cdots & B_{1r} \\
B_{21} & B_{22} & \ddots & B_{2r} \\
\vdots & \ddots & \ddots & \vdots \\
B_{r1} & B_{r2} & \cdots & B_{rr}
\end{pmatrix}
\]

Then by block multiplication, since $B$ is given to commute with $D$,

$$\lambda_j B_{ij} = \lambda_i B_{ij}$$

Therefore, if $i \neq j, B_{ij} = 0$. ■

**Lemma 14.8.8** Let $\mathcal{F}$ denote a commuting family of $n \times n$ matrices such that each $A \in \mathcal{F}$ is diagonalizable. Then $\mathcal{F}$ is simultaneously diagonalizable.

**Proof:** First note that if every matrix in $\mathcal{F}$ has only one eigenvalue, there is nothing to prove. This is because for $A$ such a matrix,

$$S^{-1}AS = \lambda I$$

and so

$$A = \lambda I$$

Thus all the matrices in $\mathcal{F}$ are diagonal matrices and you could pick any $S$ to diagonalize them all. Therefore, without loss of generality, assume some matrix in $\mathcal{F}$ has more than one eigenvalue.

The significant part of the lemma is proved by induction on $n$. If $n = 1$, there is nothing to prove because all the $1 \times 1$ matrices are already diagonal matrices. Suppose then that the theorem is true for all $k \leq n - 1$ where $n \geq 2$ and let $\mathcal{F}$ be a commuting family of diagonalizable $n \times n$ matrices. Pick $A \in \mathcal{F}$ which has more than one eigenvalue and let $S$ be an invertible matrix such that $S^{-1}AS = D$ where $D$ is of the form given in 14.13. By permuting the columns of $S$ there is no loss of generality in assuming $D$ has this form. Now denote by $\bar{\mathcal{F}}$ the collection of matrices, \(\{S^{-1}CS : C \in \mathcal{F}\}\). Note $\bar{\mathcal{F}}$ features the single matrix $S$.

It follows easily that $\bar{\mathcal{F}}$ is also a commuting family of diagonalizable matrices. By Lemma \[14.8.7\] every $B \in \bar{\mathcal{F}}$ is a block diagonal matrix of the form given in 14.8 because each of these commutes with $D$ described above as $S^{-1}AS$ and so by block multiplication, the diagonal blocks $B_i$ corresponding to different $B \in \bar{\mathcal{F}}$ commute.

By Corollary \[14.8.4\] each of these blocks is diagonalizable. This is because $B$ is known to be so. Therefore, by induction, since all the blocks are no larger than $n - 1 \times n - 1$ thanks to the assumption that $A$ has more than one eigenvalue, there exist invertible $n_i \times n_i$ matrices, $T_i$ such that $T_i^{-1}B_iT_i$ is a diagonal matrix whenever $B_i$ is one of the matrices making up the block diagonal of any $B \in \bar{\mathcal{F}}$. It follows that for $T$ defined by

\[
T = \begin{pmatrix}
T_1 & 0 & \cdots & 0 \\
0 & T_2 & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & T_r
\end{pmatrix},
\]

then $T^{-1}BT$ = a diagonal matrix for every $B \in \bar{\mathcal{F}}$ including $D$. Consider $ST$. It follows that for all $C \in \mathcal{F}$,

$$T^{-1}S^{-1}CS = T = (ST)^{-1}C(ST) = \text{ a diagonal matrix.}$$ ■

**Theorem 14.8.9** Let $\mathcal{F}$ denote a family of matrices which are diagonalizable. Then $\mathcal{F}$ is simultaneously diagonalizable if and only if $\mathcal{F}$ is a commuting family.
14.9 Fractional Powers

The main result is the following theorem.

**Theorem 14.9.1** Let $A$ be a self adjoint and nonnegative $n \times n$ matrix (all eigenvalues are nonnegative) and let $k$ be a positive integer. Then there exists a unique self adjoint nonnegative matrix $B$ such that $B^k = A$.

**Proof:** By Theorem 14.1.6, there exists an orthonormal basis of eigenvectors of $A$, say $\{v_i\}_{i=1}^n$ such that $Av_i = \lambda_i v_i$ with each $\lambda_i$ real. In particular, there exists a unitary matrix $U$ such that

$$U^*AU = D, \quad A = UDU^*$$

where $D$ has nonnegative diagonal entries. Define $B$ in the obvious way.

$$B \equiv UD^{1/k}U^*$$

Then it is clear that $B$ is self adjoint and nonnegative. Also it is clear that $B^k = A$. What of uniqueness? Let $p(t)$ be a polynomial whose graph contains the ordered pairs $\left(\lambda_i, \lambda_i^{1/k}\right)$ where the $\lambda_i$ are the diagonal entries of $D$, the eigenvalues of $A$. Then

$$p(A) = UP(D)U^* = UD^{1/k}U^* \equiv B$$

Suppose then that $C^k = A$ and $C$ is also self adjoint and nonnegative.

$$CB = CP(A) = CP(C^k) = p(C^k)C = p(A)C = BC$$

and so $\{B, C\}$ is a commuting family of non defective matrices. By Theorem 14.8.9 this family of matrices is simultaneously diagonalizable. Hence there exists a single $S$ such that

$$S^{-1}BS = D_B, \quad S^{-1}CS = D_C$$

Where $D_C, D_B$ denote diagonal matrices. Hence, raising to the power $k$, it follows that

$$A = B^k = SD_B^kS^{-1}, \quad A = C^k = SD_C^kS^{-1}$$

Hence

$$SD_B^kS^{-1} = SD_C^kS^{-1}$$

and so $D_B^k = D_C^k$. Since the entries of the two diagonal matrices are nonnegative, this implies $D_B = D_C$ and so $S^{-1}BS = S^{-1}CS$ which shows $B = C$. ■

A similar result holds for a general finite dimensional inner product space. See Problem 21 in the exercises.

14.10 Spectral Theory Of Self Adjoint Operators

First is some notation which may be useful since it will be used in the following presentation.

**Definition 14.10.1** Let $X, Y$ be inner product space and let $u, v \in X$. Then define $u \otimes v \in \mathcal{L}(X,Y)$ as follows.

$$u \otimes v(w) \equiv (w, v)u$$
where \((w, v)\) is the inner product in \(X\). Then this is clearly linear. That it is continuous follows right away from
\[
|<(w, v)u|X|w|X|v|X|
\]
and so
\[
\sup_{|w|X \leq 1} |u \otimes v (w)|Y| \leq |u|Y|v|X|
\]
Sometimes this is called the tensor product, although much more can be said about the tensor product.

Note how this is similar to the rank one transformations used to consider the dimension of the space \(\mathcal{L}(V, W)\) in Theorem 7.1.4. This is also a rank one transformation but here there is no restriction on the dimension of the vector spaces although, as usual, the interest is in finite dimensional spaces. In case you have \(\{v_1, \ldots, v_n\}\) an orthonormal basis for \(V\) and \(\{u_1, \ldots, u_m\}\) an orthonormal basis for \(Y\), (or even just a basis.) the linear transformations \(u_i \otimes v_j\) are the same as those rank one transformations used before in the above theorem and are a basis for \(\mathcal{L}(V, W)\). Thus for \(A = \sum_{i,j} a_{ij} u_i \otimes v_j\), the matrix of \(A\) with respect to the two bases has its \(ij^{th}\) entry equal to \(a_{ij}\). This is stated as the following proposition.

**Proposition 14.10.2** Suppose \(\{v_1, \ldots, v_n\}\) is an orthonormal basis for \(V\) and \(\{u_1, \ldots, u_m\}\) is a basis for \(W\). Then if \(A \in \mathcal{L}(V, W)\) is given by \(A = \sum_{i,j} a_{ij} u_i \otimes v_j\), then the matrix of \(A\) with respect to these two bases is an \(m \times n\) matrix whose \(ij^{th}\) entry is \(a_{ij}\).

In case \(A\) is a Hermitian matrix, and you have an orthonormal basis of eigenvectors and \(U\) is the unitary matrix having these eigenvectors as columns, recall that the matrix of \(A\) with respect to this basis is diagonal. Recall why this is.

**Theorem 14.10.4** Let \(A \in \mathcal{L}(X, X)\) be self adjoint (Hermitian) where \(X\) is a finite dimensional inner product space of dimension \(n\). Thus \(A = A^*\). Then there exists an orthonormal basis of eigenvectors, \(\{v_j\}_{j=1}^n\).
Proof: Consider \((Ax, x)\). This quantity is always a real number because

\[
(Ax, x) = (x, Ax) = (x, A^* x) = (Ax, x)
\]

thanks to the assumption that \(A\) is self adjoint. Now define

\[
\lambda_1 = \inf \{(Ax, x) : |x| = 1, x \in X_1 \equiv X\}.
\]

Claim: \(\lambda_1\) is finite and there exists \(v_1 \in X\) with \(|v_1| = 1\) such that \((Av_1, v_1) = \lambda_1\).

Proof of claim: Let \(\{u_j\}_{j=1}^n\) be an orthonormal basis for \(X\) and for \(x \in X\), let \((x_1, \ldots, x_n)\) be defined as the components of the vector \(x\). Thus,

\[
x = \sum_{j=1}^n x_j u_j.
\]

Since this is an orthonormal basis, it follows from the axioms of the inner product that

\[
|x|^2 = \sum_{j=1}^n |x_j|^2.
\]

Thus

\[
(Ax, x) = \left( \sum_{k=1}^n x_k Au_k, \sum_{j=1}^n x_j u_j \right) = \sum_{k,j} x_k x_j^* (Au_k, u_j),
\]

a real valued continuous function of \((x_1, \ldots, x_n)\) which is defined on the compact set

\[
K \equiv \{(x_1, \ldots, x_n) \in \mathbb{R}^n : \sum_{j=1}^n |x_j|^2 = 1\}.
\]

Therefore, it achieves its minimum from the extreme value theorem. Then define

\[
v_1 = \sum_{j=1}^n x_j u_j
\]

where \((x_1, \ldots, x_n)\) is the point of \(K\) at which the above function achieves its minimum. This proves the claim.

I claim that \(\lambda_1\) is an eigenvalue and \(v_1\) is an eigenvector. Letting \(w \in X_1 \equiv X\), the function of the real variable, \(t\), given by

\[
f(t) \equiv \frac{(A(v_1 + tw), v_1 + tw)}{|v_1 + tw|^2} = \frac{(Av_1, v_1) + 2t \text{Re}(Av_1, w) + t^2 (Aw, w)}{|v_1|^2 + 2t \text{Re}(v_1, w) + t^2 |w|^2}
\]

achieves its minimum when \(t = 0\). Therefore, the derivative of this function evaluated at \(t = 0\) must equal zero. Using the quotient rule, this implies, since \(|v_1| = 1\) that

\[
2 \text{Re}(Av_1, w) |v_1|^2 - 2 \text{Re}(v_1, w) (Av_1, v_1) = 2 \left( \text{Re}(Av_1, w) - \text{Re}(v_1, w) \lambda_1 \right) = 0.
\]

Thus \(\text{Re}(Av_1 - \lambda_1 v_1, w) = 0\) for all \(w \in X\). This implies \(Av_1 = \lambda_1 v_1\). To see this, let \(w \in X\) be arbitrary and let \(\theta\) be a complex number with \(|\theta| = 1\) and

\[
|(Av_1 - \lambda_1 v_1, w)| = \theta (Av_1 - \lambda_1 v_1, \bar{\theta} w).
\]

Then

\[
|(Av_1 - \lambda_1 v_1, w)| = \text{Re} (Av_1 - \lambda_1 v_1, \bar{\theta} w) = 0.
\]

Since this holds for all \(w\), \(Av_1 = \lambda_1 v_1\).
Continuing with the proof of the theorem, let \( X_2 \equiv \{ v_1 \} \perp \). This is a closed subspace of \( X \) and \( A : X_2 \to X_2 \) because for \( x \in X_2 \),

\[
(Ax, v_1) = (x, Av_1) = \lambda_1 (x, v_1) = 0.
\]

Let

\[
\lambda_2 = \inf \{(Ax, x) : |x| = 1, x \in X_2 \}
\]

As before, there exists \( v_2 \in X_2 \) such that \( Av_2 = \lambda_2 v_2 \), \( \lambda_1 \leq \lambda_2 \). Now let \( X_3 \equiv \{ v_1, v_2 \} \perp \) and continue in this way. As long as \( k < n \), it will be the case that \( \{ v_1, \ldots, v_k \} \perp \neq \{0\} \). This is because for \( k < n \) these vectors cannot be a spanning set and so there exists some \( w \notin \text{span}(v_1, \ldots, v_k) \). Then letting \( z \) be the closest point to \( w \) from \( \text{span}(v_1, \ldots, v_k) \), it follows that \( w - z \in \{ v_1, \ldots, v_k \} \perp \). Thus there is an decreasing sequence of eigenvalues \( \{ \lambda_k \}_{k=1}^n \) and a corresponding sequence of eigenvectors, \( \{ v_1, \ldots, v_n \} \) with this being an orthonormal set. ■

Contained in the proof of this theorem is the following important corollary.

**Corollary 14.10.5** Let \( A \in \mathcal{L}(X, X) \) be self adjoint where \( X \) is a finite dimensional inner product space. Then all the eigenvalues are real and for \( 1 \leq \lambda_2 \leq \cdots \leq \lambda_n \) the eigenvalues of \( A \), there exists an orthonormal set of vectors \( \{ u_1, \cdots, u_n \} \) for which

\[
Au_k = \lambda_k u_k.
\]

Furthermore,

\[
\lambda_k = \inf \{(Ax, x) : |x| = 1, x \in X_k \}
\]

where

\[
X_k \equiv \{ u_1, \cdots, u_{k-1} \} \perp, X_1 \equiv X.
\]

**Corollary 14.10.6** Let \( A \in \mathcal{L}(X, X) \) be self adjoint (Hermitian) where \( X \) is a finite dimensional inner product space. Then the largest eigenvalue of \( A \) is given by

\[
\max \{(Ax, x) : |x| = 1\} \tag{14.15}
\]

and the minimum eigenvalue of \( A \) is given by

\[
\min \{(Ax, x) : |x| = 1\}. \tag{14.16}
\]

**Proof:** The proof of this is just like the proof of Theorem 14.10.4. Simply replace \( \inf \) with \( \sup \) and obtain a decreasing list of eigenvalues. This establishes 14.10.4. The claim 14.10.5 follows from Theorem 14.10.4. ■

Another important observation is found in the following corollary.

**Corollary 14.10.7** Let \( A \in \mathcal{L}(X, X) \) where \( A \) is self adjoint. Then \( A = \sum_i \lambda_i v_i \otimes v_i \) where \( Av_i = \lambda_i v_i \) and \( \{ v_i \}_{i=1}^n \) is an orthonormal basis.

**Proof:** If \( v_k \) is one of the orthonormal basis vectors, \( Av_k = \lambda_k v_k \). Also,

\[
\sum_i \lambda_i v_i \otimes v_i (v_k) = \sum_i \lambda_i v_i (v_k, v_i) = \sum_i \lambda_i \delta_{ik} v_i = \lambda_k v_k.
\]

Since the two linear transformations agree on a basis, it follows they must coincide. ■

By Proposition 14.10.2 this says the matrix of \( A \) with respect to this basis \( \{ v_i \}_{i=1}^n \) is the diagonal matrix having the eigenvalues \( \lambda_1, \cdots, \lambda_n \) down the main diagonal.

The result of Courant and Fischer which follows resembles Corollary 14.10.5 but is more useful because it does not depend on a knowledge of the eigenvectors.
Theorem 14.10.8 Let \( A \in \mathcal{L}(X, X) \) be self adjoint where \( X \) is a finite dimensional inner product space. Then for \( \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \) the eigenvalues of \( A \), there exist orthonormal vectors \( \{u_1, \ldots, u_n\} \) for which
\[
Au_k = \lambda_k u_k.
\]
Furthermore,
\[
\lambda_k \equiv \max_{w_1, \ldots, w_{k-1}} \left\{ \min \left\{ \frac{(Ax, x)}{|x|} : x \neq 0, x \in \{w_1, \ldots, w_{k-1}\}^\perp \right\} \right\}
\]  \hspace{1cm} (14.17)
where if \( k = 1, \{w_1, \ldots, w_{k-1}\}^\perp \equiv X \).

**Proof:** From Theorem 14.10.3, there exist eigenvalues and eigenvectors with \( \{u_1, \ldots, u_n\} \) orthonormal and \( \lambda_i \leq \lambda_{i+1} \).

\[
(Ax, x) = \sum_{j=1}^n \lambda_j (x, u_j) (u_j, x) = \sum_{j=1}^n \lambda_j |(x, u_j)|^2
\]

Recall that \( (z, w) = \sum_j (z, u_j) (w, u_j) \). Then let \( Y = \{w_1, \ldots, w_{k-1}\}^\perp \)
\[
\inf \left\{ (Ax, x) : |x| = 1, x \in Y \right\} = \inf \left\{ \sum_{j=1}^n \lambda_j |(x, u_j)|^2 : |x| = 1, x \in Y \right\}
\]
\[
\leq \inf \left\{ \sum_{j=1}^k \lambda_j |(x, u_j)|^2 : |x| = 1, (x, u_j) = 0 \text{ for } j > k, \text{ and } x \in Y \right\}. \hspace{1cm} (14.18)
\]
The reason this is so is that the infimum is taken over a smaller set. Therefore, the infimum gets larger. Now 14.10.3 is no larger than
\[
\inf \left\{ \lambda_k \sum_{j=1}^n |(x, u_j)|^2 : |x| = 1, (x, u_j) = 0 \text{ for } j > k, \text{ and } x \in Y \right\} \leq \lambda_k
\]
because since \( \{u_1, \ldots, u_n\} \) is an orthonormal basis, \( |x|^2 = \sum_{j=1}^n |(x, u_j)|^2 \). It follows, since \( \{w_1, \ldots, w_{k-1}\} \) is arbitrary,
\[
\sup_{w_1, \ldots, w_{k-1}} \left\{ \inf \left\{ (Ax, x) : |x| = 1, x \in \{w_1, \ldots, w_{k-1}\}^\perp \right\} \right\} \leq \lambda_k. \hspace{1cm} (14.19)
\]
Then from Corollary 14.10.2, \( \lambda_k = \inf \left\{ (Ax, x) : |x| = 1, x \in \{u_1, \ldots, u_{k-1}\}^\perp \right\} \leq \sup_{w_1, \ldots, w_{k-1}} \left\{ \inf \left\{ (Ax, x) : |x| = 1, x \in \{w_1, \ldots, w_{k-1}\}^\perp \right\} \right\} \leq \lambda_k \)

Hence these are all equal and this proves the theorem. \( \blacksquare \)

The following corollary is immediate.

**Corollary 14.10.9** Let \( A \in \mathcal{L}(X, X) \) be self adjoint where \( X \) is a finite dimensional inner product space. Then for \( \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \) the eigenvalues of \( A \), there exist orthonormal vectors \( \{u_1, \ldots, u_n\} \) for which
\[
Au_k = \lambda_k u_k.
\]
Furthermore,
\[
\lambda_k \equiv \max_{w_1, \ldots, w_{k-1}} \left\{ \min \left\{ \frac{(Ax, x)}{|x|} : x \neq 0, x \in \{w_1, \ldots, w_{k-1}\}^\perp \right\} \right\}
\]  \hspace{1cm} (14.20)
where if \( k = 1, \{w_1, \ldots, w_{k-1}\}^\perp \equiv X \).
Here is a version of this for which the roles of max and min are reversed.

**Corollary 14.10.10** Let \( A \in \mathcal{L}(X, X) \) be self adjoint where \( X \) is a finite dimensional inner product space. Then for \( \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \) the eigenvalues of \( A \), there exist orthonormal vectors \( \{u_1, \cdots, u_n\} \) for which

\[
Au_k = \lambda_k u_k.
\]

Furthermore,

\[
\lambda_k = \min_{w_1, \cdots, w_{n-k}} \left\{ \max \left\{ \frac{(Ax, x)}{|x|^2} : x \neq 0, x \in \{w_1, \cdots, w_{n-k}\}^\perp \right\} \right\}
\]

(14.21)

where if \( k = n \), \( \{w_1, \cdots, w_{n-k}\}^\perp \equiv X \).

### 14.11 Positive And Negative Linear Transformations

The notion of a positive definite or negative definite linear transformation is very important in many applications. In particular it is used in versions of the second derivative test for functions of many variables. Here the main interest is the case of a linear transformation which is an \( n \times n \) matrix but the theorem is stated and proved using a more general notation because all these issues discussed here have interesting generalizations to functional analysis.

**Definition 14.11.1** A self adjoint \( A \in \mathcal{L}(X, X) \), is positive definite if whenever \( x \neq 0 \), \( (Ax, x) > 0 \) and \( A \) is negative definite if for all \( x \neq 0 \), \( (Ax, x) < 0 \). \( A \) is positive semidefinite or just nonnegative for short if for all \( x \), \( (Ax, x) \geq 0 \). \( A \) is negative semidefinite or nonpositive for short if for all \( x \), \( (Ax, x) \leq 0 \).

The following lemma is of fundamental importance in determining which linear transformations are positive or negative definite.

**Lemma 14.11.2** Let \( X \) be a finite dimensional inner product space. A self adjoint \( A \in \mathcal{L}(X, X) \) is positive definite if and only if all its eigenvalues are positive and negative definite if and only if all its eigenvalues are negative. It is positive semidefinite if all the eigenvalues are nonnegative and it is negative semidefinite if all the eigenvalues are nonpositive.

**Proof:** Suppose first that \( A \) is positive definite and let \( \lambda \) be an eigenvalue. Then for \( x \) an eigenvector corresponding to \( \lambda \), \( \lambda (x, x) = (Ax, x) = (Ax, x) > 0 \). Therefore, \( \lambda > 0 \) as claimed.

Now suppose all the eigenvalues of \( A \) are positive. From Theorem [14.10.7] and Corollary [14.10.7], \( A = \sum_{i=1}^n \lambda_i u_i \otimes u_i \) where the \( \lambda_i \) are the positive eigenvalues and \( \{u_i\} \) are an orthonormal set of eigenvectors. Therefore, letting \( x \neq 0 \),

\[
(Ax, x) = \left( \sum_{i=1}^n \lambda_i u_i \otimes u_i \right) x, x = \left( \sum_{i=1}^n \lambda_i (x, u_i), x \right) = \left( \sum_{i=1}^n \lambda_i (x, u_i), x \right) \quad \text{since \( \{u_i\} \) is an orthonormal basis, \( |x|^2 = \sum_{i=1}^n |(u_i, x)|^2 \).}
\]

because, since \( \{u_i\} \) is an orthonormal basis, \( |x|^2 = \sum_{i=1}^n |(u_i, x)|^2 \).

To establish the claim about negative definite, it suffices to note that \( A \) is negative definite if and only if \( -A \) is positive definite and the eigenvalues of \( A \) are \( (-1) \) times the eigenvalues of \( -A \). The claims about positive semidefinite and negative semidefinite are obtained similarly. ■
Definition 14.11.3 Let $A$ be an $n \times n$ matrix. Denote by $A_k$ the $k \times k$ matrix obtained by deleting the $k+1, \cdots, n$ columns and the $k+1,\cdots, n$ rows from $A$. Thus $A_n = A$ and $A_k$ is the $k \times k$ submatrix of $A$ which occupies the upper left corner of $A$. The determinants of these submatrices are called the principle minors.

The following theorem is proved in [3]. For the sake of simplicity, we state this for real matrices since this is also where the main interest lies.

**Theorem 14.11.4** Let $A$ be a self adjoint $n \times n$ matrix. Then $A$ is positive definite if and only if $\det(A_k) > 0$ for every $k = 1, \cdots, n$.

**Proof:** This theorem is proved by induction on $n$. It is clearly true if $n = 1$. Suppose then that it is true for $n-1$ where $n \geq 2$. Since $\det(A) > 0$, it follows that all the eigenvalues are nonzero. Are they all positive? Suppose not. Then there is some even number of them which are negative, even because the product of all the eigenvalues is known to be positive, equaling $\det(A)$. Pick two, $\lambda_1$ and $\lambda_2$ and let $Au_i = \lambda_i u_i$ where $u_i \neq 0$ for $i = 1, 2$ and $(u_1, u_2) = 0$. Now if $y = \alpha_1 u_1 + \alpha_2 u_2$ is an element of span $(u_1, u_2)$, then since these are eigenvalues and $(u_1, u_2)_{\mathbb{R}^n} = 0$, a short computation shows

$$(A(\alpha_1 u_1 + \alpha_2 u_2), \alpha_1 u_1 + \alpha_2 u_2) = |\alpha_1|^2 \lambda_1 |u_1|^2 + |\alpha_2|^2 \lambda_2 |u_2|^2 < 0.$$ 

Now letting $x \in \mathbb{R}^{n-1}, x \neq 0$, the induction hypothesis implies

$$(x^T, 0) A \begin{pmatrix} x \\ 0 \end{pmatrix} = x^T A_{n-1} x = (A_{n-1} x, x) > 0.$$ 

The dimension of $\{z \in \mathbb{R}^n : z_n = 0\}$ is $n-1$ and the dimension of span $(u_1, u_2) = 2$ and so there must be some nonzero $x \in \mathbb{R}^n$ which is in both of these subspaces of $\mathbb{R}^n$. However, the first computation would require that $(Ax, x) < 0$ while the second would require that $(Ax, x) > 0$. This contradiction shows that all the eigenvalues must be positive. This proves the if part of the theorem.

To show the converse, note that, as above, $(Ax, x) = x^T Ax$. Suppose that $A$ is positive definite. Then this is equivalent to having

$$x^T Ax \geq \delta \|x\|^2.$$ 

Note that for $x \in \mathbb{R}^k$,

$$\begin{pmatrix} x^T \\ 0 \end{pmatrix} A \begin{pmatrix} x \\ 0 \end{pmatrix} = x^T A_k x \geq \delta \|x\|^2.$$ 

From Lemma 14.11.1 this implies that all the eigenvalues of $A_k$ are positive. Hence from Lemma 14.11.2 it follows that $\det(A_k) > 0$, being the product of its eigenvalues.

**Corollary 14.11.5** Let $A$ be a self adjoint $n \times n$ matrix. Then $A$ is negative definite if and only if $\det(A_k)(-1)^k > 0$ for every $k = 1, \cdots, n$.

**Proof:** This is immediate from the above theorem by noting that, as in the proof of Lemma 14.11.1, $A$ is negative definite if and only if $-A$ is positive definite. Therefore, $\det(-A_k) > 0$ for all $k = 1, \cdots, n$, is equivalent to having $A$ negative definite. However, $\det(-A_k) = (-1)^k \det(A_k)$.

### 14.12 The Singular Value Decomposition

In this section, $A$ will be an $m \times n$ matrix. To begin with, here is a simple lemma.

**Lemma 14.12.1** Let $A$ be an $m \times n$ matrix. Then $A^*A$ is self adjoint and all its eigenvalues are nonnegative.

**Proof:** It is obvious that $A^*A$ is self adjoint. Suppose $A^*Ax = \lambda x$. Then $\lambda \|x\|^2 = (\lambda x, x) = (A^*Ax, x) = (Ax, Ax) \geq 0$.
**Definition 14.12.2** Let $A$ be an $m \times n$ matrix. The singular values of $A$ are the square roots of the positive eigenvalues of $A^*A$.

With this definition and lemma here is the main theorem on the singular value decomposition. In all that follows, I will write the following partitioned matrix

$$
\begin{pmatrix}
\sigma & 0 \\
0 & 0
\end{pmatrix}
$$

where $\sigma$ denotes an $r \times r$ diagonal matrix of the form

$$
\begin{pmatrix}
\sigma_1 & 0 \\
0 & \ddots \\
0 & 0 & \sigma_k
\end{pmatrix}
$$

and the bottom row of zero matrices in the partitioned matrix, as well as the right columns of zero matrices are each of the right size so that the resulting matrix is $m \times n$. Either could vanish completely. However, I will write it in the above form. It is easy to make the necessary adjustments in the other two cases.

**Theorem 14.12.3** Let $A$ be an $m \times n$ matrix. Then there exist unitary matrices, $U$ and $V$ of the appropriate size such that

$$
U^*AV = \begin{pmatrix}
\sigma & 0 \\
0 & 0
\end{pmatrix}
$$

where $\sigma$ is of the form

$$
\sigma = \begin{pmatrix}
\sigma_1 & 0 \\
0 & \ddots \\
0 & 0 & \sigma_k
\end{pmatrix}
$$

for the $\sigma_i$ the singular values of $A$, arranged in order of decreasing size.

**Proof:** By the above lemma and Theorem 14.10.4 there exists an orthonormal basis, $\{v_i\}_{i=1}^n$ for $\mathbb{F}^n$ such that $A^*Av_i = \sigma_i^2v_i$ where $\sigma_i^2 > 0$ for $i = 1, \ldots, k, (\sigma_i > 0)$, and equals zero if $i > k$. Let the eigenvalues $\sigma_i^2$ be arranged in decreasing order. It is desired to have

$$
AV = U \begin{pmatrix}
\sigma & 0 \\
0 & 0
\end{pmatrix}
$$

and so if $U = \begin{pmatrix}
u_1 & \cdots & u_m\end{pmatrix}$, one needs to have for $j \leq k$, $\sigma_j u_j = Av_j$. Thus let

$$
u_j = \sigma_j^{-1}Av_j, \quad j \leq k$$

Then for $i, j \leq k$,

$$
(u_i, u_j) = \sigma_j^{-1}\sigma_i^{-1} (Av_i, Av_j) = \sigma_j^{-1}\sigma_i^{-1} (A^*Av_i, v_j) = \sigma_j^{-1}\sigma_i^{-1}\sigma_i^2(v_i, v_j) = \delta_{ij}
$$

Now extend to an orthonormal basis of $\mathbb{F}^m$, $\{u_1, \ldots, u_k, u_{k+1}, \ldots, u_m\}$. If $i > k$,

$$
(Av_i, Av_i) = (A^*Av_i, v_i) = 0 (v_i, v_i) = 0
$$

so $Av_i = 0$. Then for $\sigma$ given as above in the statement of the theorem, it follows that

$$
AV = U \begin{pmatrix}
\sigma & 0 \\
0 & 0
\end{pmatrix}, \quad U^*AV = \begin{pmatrix}
\sigma & 0 \\
0 & 0
\end{pmatrix}
$$

The singular value decomposition has as an immediate corollary the following interesting result.
Corollary 14.12.4 Let $A$ be an $m \times n$ matrix. Then the rank of both $A$ and $A^*$ equals the number of singular values.

**Proof:** Since $V$ and $U$ are unitary, they are each one to one and onto and so it follows that

$$\text{rank } (A) = \text{rank } (U^*AV) = \text{rank } \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} = \text{number of singular values.}$$

Also since $U, V$ are unitary,

$$\text{rank } (A^*) = \text{rank } (V^*A^*U) = \text{rank } ((U^*AV)^*)$$

$$= \text{rank } \left( \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}^* \right) = \text{number of singular values.}$$

14.13 Approximation In The Frobenius Norm

The Frobenius norm is one of many norms for a matrix. It is arguably the most obvious of all norms. Here is its definition.

**Definition 14.13.1** Let $A$ be a complex $m \times n$ matrix. Then

$$\|A\|_F \equiv (\text{trace } (AA^*))^{1/2}$$

Also this norm comes from the inner product

$$(A, B)_F \equiv \text{trace } (AB^*)$$

Thus $\|A\|_F^2$ is easily seen to equal $\sum_{ij} |a_{ij}|^2$ so essentially, it treats the matrix as a vector in $\mathbb{F}^{m \times n}$.

**Lemma 14.13.2** Let $A$ be an $m \times n$ complex matrix with singular matrix

$$\Sigma = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

with $\sigma$ as defined above, $U^*AV = \Sigma$. Then

$$\|\Sigma\|_F = \|A\|_F^2$$

(14.22)

and the following hold for the Frobenius norm. If $U, V$ are unitary and of the right size,

$$\|UA\|_F = \|A\|_F, \ |UA^*V\|_F = \|A\|_F.$$

(14.23)

**Proof:** From the definition and letting $U, V$ be unitary and of the right size,

$$\|UA\|_F^2 \equiv \text{trace } (UAA^*U^*) = \text{trace } (U^*UA^*A) = \text{trace } (AA^*) = \|A\|_F^2$$

Also,

$$\|AV\|_F^2 \equiv \text{trace } (AVV^*A^*) = \text{trace } (AA^*) = \|A\|_F^2.$$

It follows

$$\|\Sigma\|_F^2 = \|U^*AV\|_F^2 = \|AV\|_F^2 = \|A\|_F^2.$$

(14.22)

Of course, this shows that

$$\|A\|_F^2 = \sum_{i} \sigma_i^2,$$

the sum of the squares of the singular values of $A$. 
Why is the singular value decomposition important? It implies

\[ A = U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} V^* \]

where \( \sigma \) is the diagonal matrix having the singular values down the diagonal. Now sometimes \( A \) is a huge matrix, 1000\( \times \)2000 or something like that. This happens in applications to situations where the entries of \( A \) describe a picture. What also happens is that most of the singular values are very small. What if you deleted those which were very small, say for all \( i \geq l \) and got a new matrix

\[ A' = U \begin{pmatrix} \sigma' & 0 \\ 0 & 0 \end{pmatrix} V^*? \]

Then the entries of \( A' \) would end up being close to the entries of \( A \) but there is much less information to keep track of. This turns out to be very useful. More precisely, letting

\[
\sigma = \begin{pmatrix} \sigma_1 & 0 \\ \vdots & \ddots \\ 0 & \sigma_r \end{pmatrix}, \quad U^* AV = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix},
\]

\[
||A - A'||_F^2 = \left| \left| U \begin{pmatrix} \sigma - \sigma' & 0 \\ 0 & 0 \end{pmatrix} V^* \right| \right|_F^2 = \sum_{k=l+1}^{r} \sigma_k^2
\]

Thus \( A \) is approximated by \( A' \) where \( A' \) has rank \( l < r \). In fact, it is also true that out of all matrices of rank \( l \), this \( A' \) is the one which is closest to \( A \) in the Frobenius norm. Here is why.

Let \( B \) be a matrix which has rank \( l \). Then from Lemma

\[
||A - B||_F^2 = \left| \left| U^* (A - B) V \right| \right|_F^2 = \left| \left| U^* AV - U^* BV \right| \right|_F^2 = \left| \left| \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} - U^* BV \right| \right|_F^2
\]

How can you make the last entry small? Clearly you should have the off diagonal entries of \( U^* BV \) equal to zero since otherwise, you could set them equal to zero and make the expression smaller. Since the singular values of \( A \) decrease from the upper left to the lower right, it follows that for \( B \) to be closest as possible to \( A \) in the Frobenius norm,

\[ U^* BV = \begin{pmatrix} \sigma' & 0 \\ 0 & 0 \end{pmatrix} \]

where the singular values \( \sigma_k \) for \( k > r \) are set equal to zero to obtain \( \sigma' \). This implies \( B = A' \) above.

This is really obvious if you look at a simple example. Say

\[
\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}
\]

for example. Then what rank 1 matrix would be closest to this one in the Frobenius norm? Obviously

\[
\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}
\]
14.14 Least Squares And Singular Value Decomposition

The singular value decomposition also has a very interesting connection to the problem of least squares solutions. Recall that it was desired to find $x$ such that $|Ax - y|$ is as small as possible. Lemma 13.3.1 shows that there is a solution to this problem which can be found by solving the system $A^*Ax = A^*y$. Each $x$ which solves this system solves the minimization problem as was shown in the lemma just mentioned. Now consider this equation for the solutions of the minimization problem in terms of the singular value decomposition.

$$A^*z = V \begin{pmatrix} \sigma & 0 \\ 0 & 0 & 0 \end{pmatrix} U^*y.$$  

Therefore, this yields the following upon using block multiplication and multiplying on the left by $V^*$.

$$\begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix} V^*x = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} U^*y. \quad (14.24)$$

One solution to this equation which is very easy to spot is

$$x = V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*y. \quad (14.25)$$

14.15 The Moore Penrose Inverse

The particular solution of the least squares problem given in (14.25) is important enough that it motivates the following definition.

**Definition 14.15.1** Let $A$ be an $m \times n$ matrix. Then the Moore Penrose inverse of $A$, denoted by $A^+$, is defined as

$$A^+ \equiv V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*.$$

Here

$$U^*AV = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

as above.

Thus $A^+y$ is a solution to the minimization problem to find $x$ which minimizes $|Ax - y|$. In fact, one can say more about this. In the following picture $M_y$ denotes the set of least squares solutions $x$ such that $A^*Ax = A^*y$.

Then $A^+(y)$ is as given in the picture.
Proposition 14.15.2 \( A^+ y \) is the solution to the problem of minimizing \(|Ax - y|\) for all \( x \) which has smallest norm. Thus
\[
|AA^+ y - y| \leq |Ax - y| \text{ for all } x
\]
and if \( x_1 \) satisfies \(|Ax_1 - y| \leq |Ax - y| \text{ for all } x\), then \(|A^+ y| \leq |x_1|\).

Proof: Consider \( x \) satisfying (14.24), equivalently \( A^* A x = A^* y \),
\[
\begin{pmatrix}
\sigma^2 & 0 \\
0 & 0
\end{pmatrix} V^* x = \begin{pmatrix}
\sigma \\
0
\end{pmatrix} U^* y
\]
which has smallest norm. This is equivalent to making \(|V^* x|\) as small as possible because \( V^* \) is unitary and so it preserves norms. For \( z \) a vector, denote by \((z)_k\) the vector in \( \mathbb{F}^k \) which consists of the first \( k \) entries of \( z \). Then if \( x \) is a solution to (14.24),
\[
\begin{pmatrix}
\sigma^2 (V^* x)_k \\
0
\end{pmatrix} = \begin{pmatrix}
\sigma (U^* y)_k \\
0
\end{pmatrix}
\]
and so \((V^* x)_k = \sigma^{-1} (U^* y)_k\). Thus the first \( k \) entries of \( V^* x \) are determined. In order to make \(|V^* x|\) as small as possible, the remaining \( n - k \) entries should equal zero. Therefore,
\[
V^* x = \begin{pmatrix}
(V^* x)_k \\
0
\end{pmatrix} = \begin{pmatrix}
\sigma^{-1} (U^* y)_k \\
0
\end{pmatrix} = \begin{pmatrix}
\sigma^{-1} \\
0
\end{pmatrix} U^* y
\]
and so
\[
x = V \begin{pmatrix}
\sigma^{-1} & 0 \\
0 & 0
\end{pmatrix} U^* y \equiv A^+ y \quad \blacksquare
\]

Lemma 14.15.3 The matrix \( A^+ \) satisfies the following conditions.
\[
AA^+ A = A, \quad A^+ AA^+ = A^+, \quad A^+ A \text{ and } AA^+ \text{ are Hermitian.} \tag{14.26}
\]

Proof: This is routine. Recall
\[
A = U \begin{pmatrix}
\sigma & 0 \\
0 & 0
\end{pmatrix} V^*
\]
and
\[
A^+ = V \begin{pmatrix}
\sigma^{-1} & 0 \\
0 & 0
\end{pmatrix} U^*
\]
so you just plug in and verify it works. \( \blacksquare \)

A much more interesting observation is that \( A^+ \) is characterized as being the unique matrix which satisfies (14.26). This is the content of the following Theorem. The conditions are sometimes called the Penrose conditions.

Theorem 14.15.4 Let \( A \) be an \( m \times n \) matrix. Then a matrix \( A_0 \), is the Moore Penrose inverse of \( A \) if and only if \( A_0 \) satisfies
\[
AA_0 A = A, \quad A_0 AA_0 = A_0, \quad A_0 A \text{ and } AA_0 \text{ are Hermitian.} \tag{14.27}
\]

Proof: From the above lemma, the Moore Penrose inverse satisfies (14.27). Suppose then that \( A_0 \) satisfies (14.27). It is necessary to verify that \( A_0 = A^+ \). Recall that from the singular value decomposition, there exist unitary matrices, \( U \) and \( V \) such that
\[
U^* A V = \Sigma \equiv \begin{pmatrix}
\sigma & 0 \\
0 & 0
\end{pmatrix}, \quad A = U \Sigma V^*.
\]
Recall that
\[ A^+ = V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^* \]

Let
\[ A_0 = V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^* \]

where \( P \) is \( r \times r \), the same size as the diagonal matrix composed of the singular values on the main diagonal.

Next use the first equation of (14.27) to write
\[
\begin{align*}
A = \overbrace{U \Sigma V^* V}^{A_0} U \Sigma V^* = U \Sigma V^* & = U \Sigma V^*. \\
\end{align*}
\]

Then multiplying both sides on the left by \( U^* \) and on the right by \( V \),
\[
\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P & Q \\ R & S \end{pmatrix} \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sigma P \sigma & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}
\]

Therefore, \( P = \sigma^{-1} \). From the requirement that \( AA_0 \) is Hermitian,
\[
\begin{align*}
A = \overbrace{U \Sigma V^* V}^{A_0} U \Sigma V^* & = U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^* \\
\end{align*}
\]

must be Hermitian. Therefore, it is necessary that
\[
\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P & Q \\ R & S \end{pmatrix} = \begin{pmatrix} \sigma P & \sigma Q \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I & \sigma Q \\ 0 & 0 \end{pmatrix}
\]
is Hermitian. Then
\[
\begin{pmatrix} I & \sigma Q \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ Q^* \sigma & 0 \end{pmatrix}
\]
and so \( Q = 0 \).

Next,
\[
\begin{align*}
V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^* U \Sigma V^* & = V \begin{pmatrix} P \sigma & 0 \\ R \sigma & 0 \end{pmatrix} V^* = V \begin{pmatrix} I & 0 \\ R \sigma & 0 \end{pmatrix} V^* \\
\end{align*}
\]
is Hermitian. Therefore, also
\[
\begin{pmatrix} I & 0 \\ R \sigma & 0 \end{pmatrix}
\]
is Hermitian. Thus \( R = 0 \) because
\[
\begin{pmatrix} I & 0 \\ R \sigma & 0 \end{pmatrix}^* = \begin{pmatrix} I & \sigma^* R^* \\ 0 & 0 \end{pmatrix}
\]
which requires \( R \sigma = 0 \). Now multiply on right by \( \sigma^{-1} \) to find that \( R = 0 \).
Use (14.28) and the second equation of (14.27) to write

\[
\begin{bmatrix}
A_0 \\
V \begin{pmatrix}
P & Q \\
R & S
\end{pmatrix} U^* U \Sigma V^* V
\end{bmatrix}
= \begin{bmatrix}
A_0 \\
V \begin{pmatrix}
P & Q \\
R & S
\end{pmatrix}
\end{bmatrix} U^*.
\]

which implies

\[
\begin{bmatrix}
P & Q \\
R & S
\end{bmatrix}
\begin{bmatrix}
\sigma^{-1} & 0 \\
0 & S
\end{bmatrix}
\begin{bmatrix}
\sigma & 0 \\
0 & 0
\end{bmatrix}
= \begin{bmatrix}
P & Q \\
R & S
\end{bmatrix}.
\]

This yields from the above in which it was shown that \(R, Q\) are both 0

\[
\begin{bmatrix}
P & Q \\
R & S
\end{bmatrix}
\begin{bmatrix}
\sigma^{-1} & 0 \\
0 & S
\end{bmatrix}
\begin{bmatrix}
\sigma^{-1} & 0 \\
0 & S
\end{bmatrix}
= \begin{bmatrix}
P & Q \\
R & S
\end{bmatrix}.
\]

Therefore, \(S = 0\) also and so

\[
V^* A_0 U \equiv \begin{bmatrix}
P & Q \\
R & S
\end{bmatrix}
= \begin{bmatrix}
\sigma^{-1} & 0 \\
0 & 0
\end{bmatrix}
\]

which says

\[
A_0 = V \begin{bmatrix}
\sigma^{-1} & 0 \\
0 & 0
\end{bmatrix} U^* \equiv A^+. \quad \blacksquare
\]

The theorem is significant because there is no mention of eigenvalues or eigenvectors in the characterization of the Moore Penrose inverse given in (14.27). It also shows immediately that the Moore Penrose inverse is a generalization of the usual inverse. See Problem 14.

14.16 The Spectral Norm And The Operator Norm

Another way of describing a norm for an \(n \times n\) matrix is as follows.

**Definition 14.16.1** Let \(A\) be an \(m \times n\) matrix. Define the spectral norm of \(A\), written as \(\|A\|_2\) to be

\[
\max \left\{ \lambda^{1/2} : \lambda \text{ is an eigenvalue of } A^* A \right\}.
\]

That is, the largest singular value of \(A\). (Note the eigenvalues of \(A^* A\) are all positive because if \(A^* A x = \lambda x\), then

\[
\lambda |x|^2 = \lambda (x, x) = (A^* A x, x) = (A x, A x) \geq 0.
\]

Actually, this is nothing new. It turns out that \(\|\cdot\|_2\) is nothing more than the operator norm for \(A\) taken with respect to the usual Euclidean norm,

\[
|x| = \left( \sum_{k=1}^n |x_k|^2 \right)^{1/2}.
\]

**Proposition 14.16.2** The following holds.

\[
\|A\|_2 = \sup \{ |A x| : |x| = 1 \} \equiv \|A\|.
\]
14.17. THE POSITIVE PART OF A HERMITIAN MATRIX

Proof: Note that $A^*A$ is Hermitian and so by Corollary 14.10.6,

$$\|A\|_2 = \max \{(A^*Ax, x)^{1/2} : |x| = 1\} = \max \{(Ax, Ax)^{1/2} : |x| = 1\}$$

$$= \max \{|Ax| : |x| = 1\} = \|A\|.$$

Here is another proof of this proposition. Recall there are unitary matrices of the right size $U, V$ such that $A = U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} V^*$ where the matrix on the inside is as described in the section on the singular value decomposition. Then since unitary matrices preserve norms,

$$\|A\| = \sup_{|x| \leq 1} \left| U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} V^*x \right| = \sup_{|V^*x| \leq 1} \left| U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} V^*x \right|$$

$$= \sup_{|y| \leq 1} \left| U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} y \right| = \sup_{|y| \leq 1} \left| \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} y \right| = \sigma_1 \equiv \|A\|_2$$

This completes the alternate proof.

From now on, $\|A\|_2$ will mean either the operator norm of $A$ taken with respect to the usual Euclidean norm or the largest singular value of $A$; whichever is most convenient.

14.17 The Positive Part Of A Hermitian Matrix

Actually, some of the most interesting functions of matrices do not come as a power series expanded about 0 which was presented earlier. One example of this situation has already been encountered in the proof of the right polar decomposition with the square root of an Hermitian transformation which had all nonnegative eigenvalues. Another example is that of taking the positive part of an Hermitian matrix. This is important in some physical models where something may depend on the positive part of the strain which is a symmetric real matrix. Obviously there is no way to consider this as a power series expanded about 0 because the function $f(r) = r^+$ is not even differentiable at 0. Therefore, a totally different approach must be considered. First the notion of a positive part is defined.

Definition 14.17.1 Let $A$ be an Hermitian matrix. Thus it suffices to consider $A$ as an element of $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ according to the usual notion of matrix multiplication. Then there exists an orthonormal basis of eigenvectors, $\{u_1, \cdots, u_n\}$ such that

$$A = \sum_{j=1}^n \lambda_j u_j \otimes u_j,$$

for $\lambda_j$ the eigenvalues of $A$, all real. Define

$$A^+ = \sum_{j=1}^n \lambda_j^+ u_j \otimes u_j$$

where $\lambda^+ \equiv \frac{|\lambda_j| + \lambda_j}{2}$.

This gives us a nice definition of what is meant but it turns out to be very important in the applications to determine how this function depends on the choice of symmetric matrix $A$. The following addresses this question.

Theorem 14.17.2 If $A, B$ be Hermitian matrices, then for $|\cdot|$ the Frobenius norm,

$$|A^+ - B^+| \leq |A - B|.$$
Proof: Let \( A = \sum_i \lambda_i v_i \otimes v_i \) and let \( B = \sum_j \mu_j w_j \otimes w_j \) where \( \{v_i\} \) and \( \{w_j\} \) are orthonormal bases of eigenvectors.

\[
|A^+ - B^+|^2 = \text{trace} \left( \sum_i \lambda_i^+ v_i \otimes v_i - \sum_j \mu_j^+ w_j \otimes w_j \right)^2 = \\
\text{trace} \left[ \sum_i (\lambda_i^+)^2 v_i \otimes v_i + \sum_j (\mu_j^+)^2 w_j \otimes w_j \\
- \sum_{i,j} \lambda_i^+ \mu_j^+ (v_i, w_j) v_i \otimes w_j - \sum_{i,j} \lambda_i^+ \mu_j^+ (v_i, w_j) w_j \otimes v_i \right]
\]

Since the trace of \( v_i \otimes w_j \) is \((v_i, w_j)\), a fact which follows from \((v_i, w_j)\) being the only possibly nonzero eigenvalue,

\[
= \sum_i (\lambda_i^+)^2 + \sum_j (\mu_j^+)^2 - 2 \sum_{i,j} \lambda_i^+ \mu_j^+ |(v_i, w_j)|^2.
\]

(14.32)

Since these are orthonormal bases,

\[
\sum_i |(v_i, w_j)|^2 = 1 = \sum_j |(v_i, w_j)|^2
\]

and so \(14.32\) equals

\[
\sum_i \sum_j \left( (\lambda_i^+)^2 + (\mu_j^+)^2 - 2 \lambda_i^+ \mu_j^+ \right) |(v_i, w_j)|^2.
\]

Similarly,

\[
|A - B|^2 = \sum_i \sum_j \left( (\lambda_i)^2 + (\mu_j)^2 - 2 \lambda_i \mu_j \right) |(v_i, w_j)|^2.
\]

Now it is easy to check that \((\lambda_i)^2 + (\mu_j)^2 - 2 \lambda_i \mu_j \geq (\lambda_i^+)^2 + (\mu_j^+)^2 - 2 \lambda_i^+ \mu_j^+\).

14.18 Exercises

1. Show \((A^*)^* = A\) and \((AB)^* = B^* A^*\).

2. Prove Corollary 14.10.10.

3. Show that if \( A \) is an \( n \times n \) matrix which has an inverse then \( A^+ = A^{-1} \).

4. Using the singular value decomposition, show that for any square matrix \( A \), it follows that \( A^* A \) is unitarily similar to \( AA^* \).

5. Let \( A, B \) be a \( m \times n \) matrices. Define an inner product on the set of \( m \times n \) matrices by

\[
(A, B)_F \equiv \text{trace} \,(AB^*).
\]

Show this is an inner product satisfying all the inner product axioms. Recall for \( M \) an \( n \times n \) matrix, \( \text{trace}(M) \equiv \sum_{i=1}^n M_{ii} \). The resulting norm, \( ||\cdot||_F \) is called the Frobenius norm and it can be used to measure the distance between two matrices.

6. It was shown that a matrix \( A \) is normal if and only if it is unitarily similar to a diagonal matrix. It was also shown that if a matrix is Hermitian, then it is unitarily similar to a real diagonal matrix. Show the converse of this last statement is also true. If a matrix is unitarily to a real diagonal matrix, then it is Hermitian.
7. Let $A$ be an $m \times n$ matrix. Show $||A||^2_F = (A, A)_F = \sum_j \sigma_j^2$ where the $\sigma_j$ are the singular values of $A$.

8. If $A$ is a general $n \times n$ matrix having possibly repeated eigenvalues, show there is a sequence $\{A_k\}$ of $n \times n$ matrices having distinct eigenvalues which has the property that the $ij^{th}$ entry of $A_k$ converges to the $ij^{th}$ entry of $A$ for all $ij$. Hint: Use Schur’s theorem.

9. Prove the Cayley Hamilton theorem as follows. First suppose to obtain a sequence $6.5.1$. Show that an $A$ is a general $n \times n$ matrix having possibly repeated eigenvalues, show there is a sequence $\{A_k\}$ of $n \times n$ matrices having distinct eigenvalues which has the property that the $ij^{th}$ entry of $A_k$ converges to the $ij^{th}$ entry of $A$ for all $ij$. Hint: Use Schur’s theorem.

10. Show directly that if $A$ is an $n \times n$ matrix and $A = A^*$ ($A$ is Hermitian) then all the eigenvalues are real and eigenvectors can be assumed to be real and that eigenvectors associated with distinct eigenvalues are orthogonal, (their inner product is zero).

11. Let $v_1, \cdots, v_n$ be an orthonormal basis for $\mathbb{F}^n$. Let $Q$ be a matrix whose $i^{th}$ column is $v_i$. Show

$$Q^* Q = QQ^* = I.$$ 

12. Show that an $n \times n$ matrix $Q$ is unitary if and only if it preserves distances. This means $|Qv| = |v|$. This was done in the text but you should try to do it for yourself.

13. Suppose $\{v_1, \cdots, v_n\}$ and $\{w_1, \cdots, w_n\}$ are two orthonormal bases for $\mathbb{F}^n$ and suppose $Q$ is an $n \times n$ matrix satisfying $Qv_i = w_i$. Then show $Q$ is unitary. If $|v| = 1$, show there is a unitary transformation which maps $v$ to $e_1$. This is done in the text but do it yourself with all details.

14. Let $A$ be a Hermitian matrix so $A = A^*$ and suppose all eigenvalues of $A$ are larger than $\delta^2$. Show

$$(A v, v) \geq \delta^2 |v|^2$$

Where here, the inner product is $(v, u) \equiv \sum_{j=1}^{n} v_j \overline{u_j}$.

15. The discrete Fourier transform maps $\mathbb{C}^n \rightarrow \mathbb{C}^n$ as follows.

$$F(x) = z \text{ where } z_k = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} e^{-i \frac{2\pi}{n} jk} x_j.$$ 

Show that $F^{-1}$ exists and is given by the formula

$$F^{-1}(z) = x \text{ where } x_j = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} e^{i \frac{2\pi}{n} jk} z_k.$$ 

Here is one way to approach this problem. Note $z = Ux$ where

$$U = \frac{1}{\sqrt{n}} 
\begin{pmatrix}
 e^{-i \frac{2\pi}{n} \cdot 0} & e^{-i \frac{2\pi}{n} \cdot 1} & e^{-i \frac{2\pi}{n} \cdot 2} & \cdots & e^{-i \frac{2\pi}{n} \cdot (n-1)} \\
 e^{-i \frac{2\pi}{n} \cdot 0} & e^{-i \frac{2\pi}{n} \cdot 1} & \cdots & e^{-i \frac{2\pi}{n} \cdot (n-1)} \\
 e^{-i \frac{2\pi}{n} \cdot 0} & e^{-i \frac{2\pi}{n} \cdot 1} & \cdots & e^{-i \frac{2\pi}{n} \cdot (n-1)} \\
 \vdots & \vdots & \ddots & \ddots \\
 e^{-i \frac{2\pi}{n} \cdot (n-1)} & e^{-i \frac{2\pi}{n} \cdot (n-1)} & \cdots & e^{-i \frac{2\pi}{n} \cdot (n-1)}
\end{pmatrix}$$

Now argue $U$ is unitary and use this to establish the result. To show this verify each row has length 1 and the inner product of two different rows gives 0. Now $U_{kj} = e^{-i \frac{2\pi}{n} jk}$ and so $(U^*)_{kj} = e^{i \frac{2\pi}{n} jk}$. 

16. Let $f$ be a periodic function having period $2\pi$. The Fourier series of $f$ is an expression of the form

$$\sum_{k=-\infty}^{\infty} c_k e^{ikx} \equiv \lim_{n \to \infty} \sum_{k=-n}^{n} c_k e^{ikx}$$

and the idea is to find $c_k$ such that the above sequence converges in some way to $f$. If

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx}$$

and you formally multiply both sides by $e^{-imx}$ and then integrate from $0$ to $2\pi$, interchanging the integral with the sum without any concern for whether this makes sense, show it is reasonable from this to expect

$$c_m = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-imx} dx.$$ 

Now suppose you only know $f(x)$ at equally spaced points $2\pi j/n$ for $j = 0, 1, \ldots, n$. Consider the Riemann sum for this integral obtained from using the left endpoint of the subintervals determined from the partition $\left\{ \frac{2\pi j}{n} \right\}_{j=0}^{n}$. How does this compare with the discrete Fourier transform? What happens as $n \to \infty$ to this approximation?

17. Suppose $A$ is a real $3 \times 3$ orthogonal matrix (Recall this means $AA^T = A^T A = I$.) having determinant 1. Show it must have an eigenvalue equal to 1. Note this shows there exists a vector $x \neq 0$ such that $Ax = x$. Hint: Show first or recall that any orthogonal matrix must preserve lengths. That is, $|Ax| = |x|$.

18. Let $A$ be a complex $m \times n$ matrix. Using the description of the Moore Penrose inverse in terms of the singular value decomposition, show that

$$\lim_{\delta \to 0^+} (A^* A + \delta I)^{-1} A^* = A^+$$

where the convergence happens in the Frobenius norm. Also verify, using the singular value decomposition, that the inverse exists in the above formula. Observe that this shows that the Moore Penrose inverse is unique.


20. In Theorem [14.9.1]. Show that every matrix which commutes with $A$ also commutes with $A^{1/k}$ the unique nonnegative self adjoint $k^{th}$ root.

21. Let $X$ be a finite dimensional inner product space and let $\beta = \{ u_1, \ldots, u_n \}$ be an orthonormal basis for $X$. Let $A \in \mathcal{L}(X, X)$ be self adjoint and nonnegative and let $M$ be its matrix with respect to the given orthonormal basis. Show that $M$ is nonnegative, self adjoint also. Use this to show that $A$ has a unique nonnegative self adjoint $k^{th}$ root.

22. Let $A$ be a complex $m \times n$ matrix having singular value decomposition $U^* AV = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$ as explained above, where $\sigma$ is $k \times k$. Show that

$$\ker(A) = \text{span}(Ve_{k+1}, \ldots, Ve_n),$$

the last $n - k$ columns of $V$. 
23. The principal submatrices of an \( n \times n \) matrix \( A \) are \( A_k \) where \( A_k \) consists those entries which are in the first \( k \) rows and first \( k \) columns of \( A \). Suppose \( A \) is a real symmetric matrix and that \( x \rightarrow \langle Ax, x \rangle \) is positive definite. This means that if \( x \neq 0 \), then \( \langle Ax, x \rangle > 0 \). Show that each of the principal submatrices are positive definite. \textbf{Hint:} Consider \( \begin{pmatrix} x^T & 0 \end{pmatrix} A \begin{pmatrix} x \\ 0 \end{pmatrix} \) where \( x \) consists of \( k \) entries.

24. Show that if \( A \) is a symmetric positive definite \( n \times n \) real matrix, then \( A \) has an \( LU \) factorization with the property that each entry on the main diagonal in \( U \) is positive. \textbf{Hint:} This is pretty clear if \( A \) is \( 1 \times 1 \). Assume true for \((n-1) \times (n-1)\). Then

\[
A = \begin{pmatrix} \hat{A} & a \\ a^T & a_{nn} \end{pmatrix}
\]

Then as above, \( \hat{A} \) is positive definite. Thus it has an \( LU \) factorization with all positive entries on the diagonal of \( U \). Notice that, using block multiplication,

\[
A = \begin{pmatrix} LU & a \\ a^T & a_{nn} \end{pmatrix} = \begin{pmatrix} L & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} U & L^{-1}a \\ a^T & a_{nn} \end{pmatrix}
\]

Now consider that matrix on the right. Argue that it is of the form \( \tilde{L} \tilde{U} \) where \( \tilde{U} \) has all positive diagonal entries except possibly for the one in the \( n^{th} \) row and \( n^{th} \) column. Now explain why \( \det(A) > 0 \) and argue that in fact all diagonal entries of \( \tilde{U} \) are positive.

25. Let \( A \) be a real symmetric \( n \times n \) matrix and \( A = LU \) where \( L \) has all ones down the diagonal and \( U \) has all positive entries down the main diagonal. Show that \( A = LDH \) where \( L \) is lower triangular and \( H \) is upper triangular, each having all ones down the diagonal and \( D \) a diagonal matrix having all positive entries down the main diagonal. In fact, these are the diagonal entries of \( U \).

26. Show that if \( L, L_1 \) are lower triangular with ones down the main diagonal and \( H, H_1 \) are upper triangular with all ones down the main diagonal and \( D, D_1 \) are diagonal matrices having all positive diagonal entries, and if \( LDH = L_1D_1H_1 \), then \( L = L_1, H = H_1, D = D_1 \). \textbf{Hint:} Explain why \( D_1^{-1}L_1^{-1}LD = H_1H^{-1} \). Then explain why the right side is upper triangular and the left side is lower triangular. Conclude these are both diagonal matrices. However, there are all ones down the diagonal in the expression on the right. Hence \( H = H_1 \). Do something similar to conclude that \( L = L_1 \) and then that \( D = D_1 \).

27. Show that if \( A \) is a symmetric real matrix such that \( x \rightarrow \langle Ax, x \rangle \) is positive definite, then there exists a lower triangular matrix \( L \) having all positive entries down the diagonal such that \( A = LL^T \). \textbf{Hint:} From the above, \( A = LDH \) where \( L, H \) are respectively lower and upper triangular having all ones down the diagonal and \( D \) is a diagonal matrix having all positive entries. Then argue from the above problem and symmetry of \( A \) that \( H = L^T \). Now modify \( L \) by making it equal to \( LD^{1/2} \). This is called the Cholesky factorization.

28. Given \( F \in L(X,Y) \) where \( X,Y \) are inner product spaces and \( \dim(X) = n \leq m = \dim(Y) \), there exists \( R, U \) such that \( U \) is nonnegative and Hermitian \( (U = U^*) \) and \( R^*R = I \) such that \( F = RU \). Show that \( U \) is actually unique and that \( R \) is determined on \( U(X) \).

29. If \( A \) is a complex Hermitian \( n \times n \) matrix which has all eigenvalues nonnegative, show that there exists a complex Hermitian matrix \( B \) such that \( BB = A \).

30. Suppose \( A, B \) are \( n \times n \) real Hermitian matrices and they both have all nonnegative eigenvalues. Show that \( \det(A + B) \geq \det(A) + \det(B) \). \textbf{Hint:} Use the above problem and the Cauchy Binet theorem. Let \( P^2 = A, Q^2 = B \) where \( P, Q \) are Hermitian and nonnegative. Then

\[
A + B = \begin{pmatrix} P & Q \end{pmatrix} \begin{pmatrix} P \\ Q \end{pmatrix}
\]
31. Suppose \( B = \begin{pmatrix} \alpha & c^* \\ b & A \end{pmatrix} \) is an \((n+1) \times (n+1)\) Hermitian nonnegative matrix where \( \alpha \) is a scalar and \( A \) is \( n \times n \). Show that \( \alpha \) must be real, \( c = b \), and \( A = A^* \), \( A \) is nonnegative, and that if \( \alpha = 0 \), then \( b = 0 \). Otherwise, \( \alpha > 0 \).

32. If \( A \) is an \( n \times n \) complex Hermitian and nonnegative matrix, show that there exists an upper triangular matrix \( B \) such that \( B^* B = A \). **Hint:** Prove this by induction. It is obviously true if \( n = 1 \). Now if you have an \((n + 1) \times (n + 1)\) Hermitian nonnegative matrix, then from the above problem, it is of the form \( \begin{pmatrix} \alpha^2 & \alpha b^* \\ \alpha b & A \end{pmatrix} \), \( \alpha \) real.

33. **Suppose** \( A \) is a nonnegative Hermitian matrix (all eigenvalues are nonnegative) which is partitioned as
\[
A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}
\]
where \( A_{11}, A_{22} \) are square matrices. Show that \( \det (A) \leq \det (A_{11}) \det (A_{22}) \). **Hint:** Use the above problem to factor \( A \) getting
\[
A = \begin{pmatrix} B_{11}^* & 0^* \\ B_{12} & B_{22}^* \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}
\]
Next argue that \( A_{11} = B_{11}^* B_{11}, A_{22} = B_{12}^* B_{12} + B_{22}^* B_{22} \). Use the Cauchy Binet theorem to argue that \( \det (A_{22}) = \det (B_{12}^* B_{12} + B_{22}^* B_{22}) \geq \det (B_{22}^* B_{22}) \). Then explain why
\[
\det (A) = \det (B_{11}^*) \det (B_{22}^*) \det (B_{11}) \det (B_{22})
\]
\[
= \det (B_{11}^* B_{11}) \det (B_{22}^* B_{22})
\]

34. **Suppose** \( A \) is a Hermitian matrix which is nonnegative (all eigenvalues are nonnegative), then \( \det (A) \leq \prod_i A_{ii} \).
Chapter 15

Analysis Of Linear Transformations

15.1 The Condition Number

Let $A \in \mathcal{L}(X, X)$ be a linear transformation where $X$ is a finite dimensional vector space and consider the problem $Ax = b$ where it is assumed there is a unique solution to this problem. How does the solution change if $A$ is changed a little bit and if $b$ is changed a little bit? This is clearly an interesting question because you often do not know $A$ and $b$ exactly. If a small change in these quantities results in a large change in the solution, $x$, then it seems clear this would be undesirable.

In what follows $||\cdot||$ when applied to a linear transformation will always refer to the operator norm. Recall the following property of the operator norm in Theorem 11.6.2.

Lemma 15.1.1 Let $A, B \in \mathcal{L}(X, X)$ where $X$ is a normed vector space as above. Then for $||\cdot||$ denoting the operator norm,

$$||AB|| \leq ||A|| ||B||.$$ (15.1)

Lemma 15.1.2 Let $A, B \in \mathcal{L}(X, X), A^{-1} \in \mathcal{L}(X, X)$, and suppose $||B|| < 1/||A^{-1}||$. Then $(A + B)^{-1}, (I + A^{-1}B)^{-1}$ exists and

$$||I + A^{-1}B||^{-1} \leq (1 - ||A^{-1}B||)^{-1}$$ (15.2)

The above formula makes sense because $||A^{-1}B|| < 1$.

Proof: By Lemma 11.6.2,

$$||A^{-1}B|| \leq ||A^{-1}|| ||B|| < ||A^{-1}|| \frac{1}{||A^{-1}||} = 1$$ (15.3)

Then from the triangle inequality,

$$||(I + A^{-1}B)x|| \geq ||x|| - ||A^{-1}Bx|| \geq ||x|| - ||A^{-1}B|| ||x|| = (1 - ||A^{-1}B||) ||x||$$

It follows that $I + A^{-1}B$ is one to one because $1 - ||A^{-1}B|| > 0$. Thus if $(I + A^{-1}B)x = 0$, then $x = 0$. Thus $I + A^{-1}B$ is also onto, taking a basis to a basis. Then a generic $y \in X$ is of the form $y = (I + A^{-1}B)x$ and the above shows that

$$||(I + A^{-1}B)^{-1}y|| \leq (1 - ||A^{-1}B||^{-1}) ||y||$$
which verifies (15.1). Thus \((A + B) = A(I + A^{-1}B)\) is one to one and this with Lemma (11.6.2) implies (15.2). □

**Proposition 15.1.3** Suppose \(A\) is invertible, \(b \neq 0\), \(Ax = b\), and \((A + B)x_1 = b_1\) where \(\|B\| < 1/\|A^{-1}\|\). Then

\[
\frac{\|x_1 - x\|}{\|x\|} \leq \frac{\|A^{-1}\| \|A\|}{1 - \|A^{-1}B\|} \left( \frac{\|b_1 - b\| + \|B\|}{\|b\|} + \|A\| \right)
\]

**Proof:** This follows from the above lemma.

\[
\frac{\|x_1 - x\|}{\|x\|} = \frac{\left\| (I + A^{-1}B)^{-1} A^{-1}b_1 - A^{-1}b \right\|}{\|A^{-1}b\|} \leq \frac{1}{1 - \|A^{-1}B\|} \frac{\|A^{-1}b_1 - (I + A^{-1}B) A^{-1}b\|}{\|A^{-1}b\|} \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}B\|} \left( \frac{\|b_1 - b\| + \|B\|}{\|A^{-1}b\|} \right)
\]

because \(A^{-1}b/\|A^{-1}b\|\) is a unit vector. Now multiply and divide by \(\|A\|\). Then

\[
\leq \frac{\|A^{-1}\| \|A\|}{1 - \|A^{-1}B\|} \left( \frac{\|b_1 - b\|}{\|b\|} + \|A\| \right)
\]

This shows that the number, \(\|A^{-1}\| \|A\|\), controls how sensitive the relative change in the solution of \(Ax = b\) is to small changes in \(A\) and \(b\). This number is called the condition number. It is bad when this number is large because a small relative change in \(b\), for example, could yield a large relative change in \(x\).

Recall that for \(A\) an \(n \times n\) matrix, \(\|A\|_2 = \sigma_1\) where \(\sigma_1\) is the largest singular value. The largest singular value of \(A^{-1}\) is therefore, \(1/\sigma_n\) where \(\sigma_n\) is the smallest singular value of \(A\). Therefore, the condition number reduces to \(\sigma_1/\sigma_n\), the ratio of the largest to the smallest singular value of \(A\) provided the norm is the usual Euclidean norm.

**15.2 The Spectral Radius**

Even though it is in general impractical to compute the Jordan form, its existence is all that is needed in order to prove an important theorem about something which is relatively easy to compute. This is the spectral radius of a matrix.

**Definition 15.2.1** Define \(\sigma(A)\) to be the eigenvalues of \(A\). Also,

\[
\rho(A) \equiv \max (|\lambda| : \lambda \in \sigma(A))
\]

The number, \(\rho(A)\) is known as the spectral radius of \(A\).

Recall the following symbols and their meaning.

\[
\limsup_{n \to \infty} a_n, \liminf_{n \to \infty} a_n
\]
They are respectively the largest and smallest limit points of the sequence \( \{a_n\} \) where \( \pm \infty \) is allowed in the case where the sequence is unbounded. They are also defined as

\[
\begin{align*}
\limsup_{n \to \infty} a_n &= \lim_{n \to \infty} \left( \sup \{a_k : k \geq n\} \right), \\
\liminf_{n \to \infty} a_n &= \lim_{n \to \infty} \left( \inf \{a_k : k \geq n\} \right).
\end{align*}
\]

Thus, the limit of the sequence exists if and only if these are both equal to the same real number. Also note that the

**Lemma 15.2.2** Let \( J \) be a \( p \times p \) Jordan matrix

\[
J = \begin{pmatrix}
J_1 & & \\
& \ddots & \\
& & J_s
\end{pmatrix}
\]

where each \( J_k \) is of the form

\[
J_k = \lambda_k I + N_k
\]

in which \( N_k \) is a nilpotent matrix having zeros down the main diagonal and ones down the super diagonal. Then

\[
\lim_{n \to \infty} \|J^n\|^{1/n} = \rho
\]

where \( \rho = \max \{ |\lambda_k|, k = 1, \ldots, n \} \). Here the norm is the operator norm.

**Proof:** Consider one of the blocks, \( |\lambda_k| < \rho \). Here \( J_k \) is \( p \times p \).

\[
\frac{1}{\rho^n} J_k^n = \frac{1}{\rho^n} \sum_{i=0}^{p} \left( \begin{array}{c} n \\ i \end{array} \right) \lambda_k^{n-i} N_k^i
\]

Then

\[
\left\| \frac{1}{\rho^n} J_k^n \right\| \leq \sum_{i=0}^{p} \left( \begin{array}{c} n \\ i \end{array} \right) \|N_k^i\| \left| \lambda_k^{n-i} \right| \frac{1}{\rho^i} \rho^n = C \sum_{i=0}^{p} \frac{1}{\rho^i} \rho^n \left| \lambda_k^{n-i} \right|
\]

Now there are \( p \) numbers \( \|N_k^i\| \) so you could pick the largest, \( C \). Also

\[
\frac{\left| \lambda_k^{n-i} \right|}{\rho^{n-i}} \leq \frac{\left| \lambda_k^{n-p} \right|}{\rho^{n-p}}
\]

so \( 15.4 \) is dominated by

\[
\leq C n^p \left| \lambda_k^{n-p} \right| \frac{1}{\rho^{n-p}} \sum_{i=0}^{p} \frac{1}{\rho^i} \rho^n \left| \lambda_k^{n-p} \right|
\]

The ratio or root test shows that this converges to 0 as \( n \to \infty \).

What happens when \( |\lambda_k| = \rho^2 \)

\[
\frac{1}{\rho^n} J_k^n = \omega^n I + \sum_{i=1}^{p} \left( \begin{array}{c} n \\ i \end{array} \right) N_k^i \omega^{n-i} \frac{1}{\rho^i}
\]

where \( |\omega| = 1 \).

\[
\frac{1}{\rho^n} \|J^n\| \leq 1 + n^p C
\]

where \( C = \max \{ \|N_k^i\|, i = 1, \ldots, p, k = 1, \ldots, s \} \sum_{i=1}^{p} \frac{1}{\rho^i} \). Thus

\[
\frac{1}{\rho^n} \|J^n\| \leq \frac{1}{\rho^n} \sum_{k=1}^{s} \|J_k^n\| \leq s (1 + n^p C) = s n^p C \left( \frac{1}{n^p C} + 1 \right)
\]
CHAPTER 15. ANALYSIS OF LINEAR TRANSFORMATIONS

and so

\[
\frac{1}{p} \lim \sup_{n \to \infty} \|J^n\|^{1/n} \leq \lim \sup_{n \to \infty} s^{1/n} (n^pC)^{1/n} \left( \frac{1}{n^pC} + 1 \right)^{1/n} = 1
\]

\[
\lim \sup_{n \to \infty} \|J^n\|^{1/n} \leq \rho
\]

Next let \( x \) be an eigenvector for \( \lambda, |\lambda| = \rho \) and let \( \|x\| = 1. \) Then

\[
\rho^n = \rho^n \|x\| = \|J^n x\| \leq \|J^n\|
\]

and so

\[
\|J^n\|^{1/n} = \rho^n
\]

Hence

\[
\rho \geq \lim \sup_{n \to \infty} \|J^n\|^{1/n} \geq \lim \inf_{n \to \infty} \|J^n\|^{1/n} \geq \rho
\]

The following theorem is due to Gelfand around 1941.

**Theorem 15.2.3** (Gelfand) Let \( A \) be a complex \( p \times p \) matrix. Then if \( \rho \) is the absolute value of its largest eigenvalue,

\[
\lim_{n \to \infty} \|A^n\|^{1/n} = \rho.
\]

Here \(||\cdot||\) is any norm on \( \mathcal{L}(\mathbb{C}^n, \mathbb{C}^n)\).

**Proof:** First assume \(||\cdot||\) is the operator norm with respect to the usual Euclidean metric on \( \mathbb{C}^n \). Then letting \( J \) denote the Jordan form of \( A, S^{-1}AS = J \), it follows from Lemma \( \text{of} \) \( \text{Lemma} \)

\[
\lim \sup_{n \to \infty} \|A^n\|^{1/n} = \lim \sup_{n \to \infty} \|SJ^nS^{-1}\|^{1/n} \leq \lim \sup_{n \to \infty} \left( \|S\| \|S^{-1}\| \|J^n\| \right)^{1/n}
\]

\[
\leq \lim \sup_{n \to \infty} \left( \|S\| \|S^{-1}\| \|J^n\| \right)^{1/n} = \rho
\]

Letting \( \lambda \) be the largest eigenvalue of \( A, |\lambda| = \rho, \) and \( Ax = \lambda x \) where \( \|x\| = 1, \)

\[
\|A^n\| \geq \|A^n x\| = \rho^n
\]

and so

\[
\lim \inf_{n \to \infty} \|A^n\|^{1/n} \geq \rho \geq \lim \sup_{n \to \infty} \|A^n\|^{1/n}
\]

If follows that \( \lim \inf_{n \to \infty} \|A^n\|^{1/n} = \lim \sup_{n \to \infty} \|A^n\|^{1/n} = \lim_{n \to \infty} \|A^n\|^{1/n} = \rho. \)

Now by equivalence of norms, if \(||\cdot||\) is any other norm for the set of complex \( p \times p \) matrices, there exist constants \( \delta, \Delta \) such that

\[
\delta \|A^n\| \leq ||A^n|| \leq \Delta \|A^n\|
\]

Then

\[
\delta^{1/n} \|A^n\|^{1/n} \leq ||A^n||^{1/n} \leq \Delta^{1/n} \|A^n\|^{1/n}
\]

The limits exist and equal \( \rho \) for the ends of the above inequality. Hence, by the squeezing theorem, \( \rho = \lim_{n \to \infty} ||A^n||^{1/n}. \)

**Example 15.2.4** Consider \( \begin{pmatrix} 9 & -1 & 2 \\ -2 & 8 & 4 \\ 1 & 1 & 8 \end{pmatrix} \). Estimate the absolute value of the largest eigenvalue.
A laborious computation reveals the eigenvalues are 5, and 10. Therefore, the right answer in this case is 10. Consider $|A^7|^{1/7}$ where the norm is obtained by taking the maximum of all the absolute values of the entries. Thus

$$
\begin{pmatrix}
  9 & -1 & 2 \\
  -2 & 8 & 4 \\
   1 & 1 & 8
\end{pmatrix}^7 = \begin{pmatrix}
  8015625 & -1984375 & 3968750 \\
  -3968750 & 6031250 & 7937500 \\
  1984375 & 1984375 & 6031250
\end{pmatrix}
$$

and taking the seventh root of the largest entry gives

$$\rho(A) \approx 8015625^{1/7} = 9.68895123671.$$

Of course the interest lies primarily in matrices for which the exact roots to the characteristic equation are not known and in the theoretical significance.

### 15.3 Series And Sequences Of Linear Operators

Before beginning this discussion, it is necessary to define what is meant by convergence in $\mathcal{L}(X,Y)$.

**Definition 15.3.1** Let $\{A_k\}_{k=1}^{\infty}$ be a sequence in $\mathcal{L}(X,Y)$ where $X,Y$ are finite dimensional normed linear spaces. Then $\lim_{n \to \infty} A_k = A$ if for every $\varepsilon > 0$ there exists $N$ such that if $n > N$, then

$$\|A - A_n\| < \varepsilon.$$

Here the norm refers to any of the norms defined on $\mathcal{L}(X,Y)$. By Corollary 11.5.3 and Theorem 5.1.4, it doesn't matter which one is used. Define the symbol for an infinite sum in the usual way. Thus

$$\sum_{k=1}^{\infty} A_k \equiv \lim_{n \to \infty} \sum_{k=1}^{n} A_k$$

**Lemma 15.3.2** Suppose $\{A_k\}_{k=1}^{\infty}$ is a sequence in $\mathcal{L}(X,Y)$ where $X,Y$ are finite dimensional normed linear spaces. Then if

$$\sum_{k=1}^{\infty} \|A_k\| < \infty,$$

it follows that

$$\sum_{k=1}^{\infty} A_k$$

exists (converges). In words, absolute convergence implies convergence. Also,

$$\left\| \sum_{k=1}^{\infty} A_k \right\| \leq \sum_{k=1}^{\infty} \|A_k\|$$

**Proof:** For $p \leq m \leq n$,

$$\left\| \sum_{k=1}^{n} A_k - \sum_{k=1}^{m} A_k \right\| \leq \sum_{k=m}^{\infty} \|A_k\|$$

and so for $p$ large enough, this term on the right in the above inequality is less than $\varepsilon$. Since $\varepsilon$ is arbitrary, this shows the partial sums of $\sum_{k=1}^{\infty} A_k$ are a Cauchy sequence. Therefore by Corollary 11.5.3, it follows that these partial sums converge. As to the last claim,

$$\left\| \sum_{k=1}^{n} A_k \right\| \leq \sum_{k=1}^{n} \|A_k\| \leq \sum_{k=1}^{\infty} \|A_k\|$$
Therefore, passing to the limit,
\[ \left\| \sum_{k=1}^{\infty} A_k \right\| \leq \sum_{k=1}^{\infty} \|A_k\|. \]

Why is this last step justified? (Recall the triangle inequality \( \|A\| - \|B\| \leq \|A - B\| \).)

Now here is a useful result for differential equations.

**Theorem 15.3.3** Let \( X \) be a finite dimensional inner product space and let \( A \in \mathcal{L}(X, X) \). Define
\[ \Phi(t) = \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} \]

Then the series converges for each \( t \in \mathbb{R} \). Also
\[ \Phi'(t) = \lim_{h \to 0} \frac{\Phi(t + h) - \Phi(t)}{h} = \sum_{k=1}^{\infty} \frac{t^{k-1} A^k}{(k-1)!} = A \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} = A \Phi(t) \]

Also \( A \Phi(t) = \Phi(t) A \) and for all \( t \), \( \Phi(t) \Phi(-t) = I \) so \( \Phi(t)^{-1} = \Phi(-t) \), \( \Phi(0) = I \). (It is understood that \( A^0 = I \) in the above formula.)

**Proof:** First consider the claim about convergence.
\[ \sum_{k=0}^{\infty} \left| \frac{t^k A^k}{k!} \right| = \sum_{k=0}^{\infty} \frac{|t|^k \|A\|^k}{k!} = e^{\|t\|\|A\|} < \infty \]

so it converges by Lemma 15.3.2. Here \( \theta_k \in (0, 1) \). Thus
\[ \frac{\Phi(t + h) - \Phi(t)}{h} = 1 \sum_{k=0}^{\infty} \left( \frac{(t + h)^k - t^k}{k!} \right) A^k \]
\[ = \frac{1}{h} \sum_{k=0}^{\infty} \left( k (t + \theta_k h)^{k-1} - 1 \right) A^k \]
\[ = \frac{1}{h} \sum_{k=0}^{\infty} \left( k (t + \theta_k h)^{k-1} \right) A^k \]
\[ = \sum_{k=0}^{\infty} \left( t + \theta_k h \right)^{k-1} A^k \frac{1}{(k-1)!} \]
\[ = \sum_{k=0}^{\infty} \left( (k-1) \left( t + \theta_k h \right)^{k-2} \theta_k h \right) A^k \frac{1}{(k-1)!} \]
\[ = \left| \frac{\sum_{k=2}^{\infty} \left( t + \theta_k h \right)^{k-2} \theta_k h \right| A^k \frac{1}{(k-1)!} \right| \]
\[ \leq |h| \sum_{k=2}^{\infty} \frac{|t| + |h|^k \|A\|^k-2}{(k-2)!} \|A\|^2 \]
Then this equals \((x, y)\) when \(t = 0\). Take its derivative.

\[
\langle -\Phi'(-t) \Phi(t) + \Phi(-t) \Phi'(t) \rangle x, y \rangle_X
\]

\[
= \langle -A \Phi(-t) \Phi(t) + \Phi(-t) A \Phi(t) \rangle x, y \rangle_X
\]

\[
= (0, y)_X = 0
\]

Hence this scalar valued function equals a constant and so the constant must be \((x, y)_X\). Hence for all \(x, y, (\Phi(-t) \Phi(t) x - x, y)_X = 0\) for all \(x, y\) and this is in particular for \(y = \Phi(-t) \Phi(t) x - x\) which shows that \(\Phi(-t) \Phi(t) = I\). ■

As a special case, suppose \(\lambda \in \mathbb{C}\) and consider

\[
\sum_{k=0}^{\infty} \frac{t^k \lambda^k}{k!}
\]

where \(t \in \mathbb{R}\). In this case, \(A_k = \frac{t^k \lambda^k}{k!}\) and you can think of it as being in \(L(\mathbb{C}, \mathbb{C})\). Then the following corollary is of great interest.

**Corollary 15.3.4** Let

\[
f(t) \equiv \sum_{k=0}^{\infty} \frac{t^k \lambda^k}{k!} = 1 + \sum_{k=1}^{\infty} \frac{t^k \lambda^k}{k!}
\]

Then this function is a well defined complex valued function and furthermore, it satisfies the initial value problem,

\[
y' = \lambda y, \quad y(0) = 1
\]

Furthermore, if \(\lambda = a + ib\),

\[
|f|(t) = e^{at}.
\]

**Proof:** The first part is a special case of the above theorem. Note that for \(f(t) = u(t) + iv(t)\), both \(u, v\) are differentiable. This is because

\[
\begin{align*}
u &= \frac{f + \overline{f}}{2}, \\
v &= \frac{f - \overline{f}}{2i}
\end{align*}
\]

Then from the differential equation,

\[
(a + ib) (u + iv) = u' + iv'
\]

and equating real and imaginary parts,

\[
u' = au - bv, \quad v' = av + bu.
\]

Then a short computation shows

\[
(u^2 + v^2)' = 2uu' + 2vv' = 2u (au - bv) + 2v (av + bu) = 2a (u^2 + v^2)
\]

\[
(u^2 + v^2)(0) = |f|^2(0) = 1
\]

Now in general, if

\[
y' = cy, \quad y(0) = 1,
\]

with \(c\) real it follows \(y(t) = e^{ct}\). To see this,

\[
y' - cy = 0
\]

and so, multiplying both sides by \(e^{-ct}\) you get

\[
\frac{d}{dt} (ye^{-ct}) = 0
\]

and so \(ye^{-ct}\) equals a constant which must be 1 because of the initial condition \(y(0) = 1\). Thus

\[
(u^2 + v^2)(t) = e^{2at}
\]

and taking square roots yields the desired conclusion. ■
Definition 15.3.5 The function in Corollary 15.3.4 given by that power series is denoted as
\[
\exp(\lambda t) \text{ or } e^{\lambda t}.
\]

The next lemma is normally discussed in advanced calculus courses but is proved here for the convenience of the reader. It is known as the root test.

Definition 15.3.6 For \( \{a_n\} \) any sequence of real numbers
\[
\limsup_{n \to \infty} a_n \equiv \lim_{n \to \infty} (\sup \{a_k : k \geq n\})
\]
Similarly
\[
\liminf_{n \to \infty} a_n \equiv \lim_{n \to \infty} (\inf \{a_k : k \geq n\})
\]
In case \( A_n \) is an increasing (decreasing) sequence which is unbounded above (below) then it is understood that \( \lim_{n \to \infty} A_n = \infty \) (\( -\infty \)) respectively. Thus either of \( \limsup \) or \( \liminf \) can equal \( +\infty \) or \( -\infty \). However, the important thing about these is that unlike the limit, these always exist.

It is convenient to think of these as the largest point which is the limit of some subsequence of \( \{a_n\} \) and the smallest point which is the limit of some subsequence of \( \{a_n\} \) respectively. Thus \( \lim_{n \to \infty} a_n \) exists and equals some point of \( [-\infty, \infty] \) if and only if the two are equal.

Lemma 15.3.7 Let \( \{a_p\} \) be a sequence of nonnegative terms and let
\[
r = \limsup_{p \to \infty} a_p^{1/p}.
\]
Then if \( r < 1 \), it follows the series, \( \sum_{k=1}^{\infty} a_k \) converges and if \( r > 1 \), then \( a_p \) fails to converge to 0 so the series diverges. If \( A \) is an \( n \times n \) matrix and
\[
r = \limsup_{p \to \infty} \|AP\|^{1/p},
\]
then if \( r > 1 \), then \( \sum_{k=0}^{\infty} A^k \) fails to converge and if \( r < 1 \) then the series converges. Note that the series converges when the spectral radius is less than one and diverges if the spectral radius is larger than one. In fact, \( \limsup_{p \to \infty} \|AP\|^{1/p} = \lim_{p \to \infty} \|AP\|^{1/p} \) from Theorem 15.3.8.

Proof: Suppose \( r < 1 \). Then there exists \( N \) such that if \( p > N \),
\[
a_p^{1/p} < R
\]
where \( r < R < 1 \). Therefore, for all such \( p \), \( a_p < R^p \) and so by comparison with the geometric series, \( \sum R^p \), it follows \( \sum_{p=1}^{\infty} a_p \) converges.

Next suppose \( r > 1 \). Then letting \( 1 < R < r \), it follows there are infinitely many values of \( p \) at which
\[
R < a_p^{1/p}
\]
which implies \( R^p < a_p \), showing that \( a_p \) cannot converge to 0 and so the series cannot converge either.

To see the last claim, if \( r > 1 \), then \( \|AP\| \) fails to converge to 0 and so \( \{\sum_{k=0}^{m} A^k\}_{m=0}^{\infty} \) is not a Cauchy sequence. Hence \( \sum_{k=0}^{\infty} A^k \equiv \lim_{m \to \infty} \sum_{k=0}^{m} A^k \) cannot exist. If \( r < 1 \), then for all \( n \) large enough, \( \|A^n\|^{1/n} \leq r < 1 \) for some \( r \) so \( \|A^n\| \leq r^n \). Hence \( \sum_{n=1}^{\infty} \|A^n\| \) converges and so by Lemma 15.3.8, it follows that \( \sum_{k=1}^{\infty} A^k \) also converges.

Now denote by \( \sigma (A)^p \) the collection of all numbers of the form \( \lambda^p \) where \( \lambda \in \sigma (A) \).

Lemma 15.3.8 \( \sigma (A)^p = \sigma (A)^p \equiv \{\lambda^p : \lambda \in \sigma (A)\} \).
**Proof:** In dealing with $\alpha (A^p)$, it suffices to deal with $\sigma (J^p)$ where $J$ is the Jordan form of $A$ because $J^p$ and $A^p$ are similar. Thus if $\lambda \in \sigma (A^p)$, then $\lambda \in \sigma (J^p)$ and so $\lambda = \alpha$ where $\alpha$ is one of the entries on the main diagonal of $J^p$. These entries are of the form $\lambda^p$ where $\lambda \in \sigma (A)$. Thus $\lambda \in \sigma (A^p)$ and this shows $\sigma (A^p) \subseteq \sigma (A)^p$.

Now take $\alpha \in \sigma (A)$ and consider $\alpha^p$.

$$\alpha^p I - A^p = (\alpha^{p-1} I + \cdots + \alpha A^{p-2} + A^{p-1}) (\alpha I - A)$$

and so $\alpha^p I - A^p$ fails to be one to one which shows that $\alpha^p \in \sigma (A^p)$ which shows that $\sigma (A)^p \subseteq \sigma (A^p)$.

### 15.4 Iterative Methods For Linear Systems

Consider the problem of solving the equation

$$A x = b$$

(15.7)

where $A$ is an $n \times n$ matrix. In many applications, the matrix $A$ is huge and composed mainly of zeros. For such matrices, the method of Gauss elimination (row operations) is not a good way to solve the system because the row operations can destroy the zeros and storing all those zeros takes a lot of room in a computer. These systems are called sparse. To solve them, it is common to use an iterative technique. I am following the treatment given to this subject by Nobel and Daniel [27].

**Definition 15.4.1** The Jacobi iterative technique, also called the method of simultaneous corrections is defined as follows. Let $x^1$ be an initial vector, say the zero vector or some other vector. The method generates a succession of vectors, $x^2, x^3, x^4, \ldots$ and hopefully this sequence of vectors will converge to the solution to (15.7). The vectors in this list are called iterates and they are obtained according to the following procedure. Letting $A = (a_{ij})$,

$$a_{ii} x_i^{r+1} = - \sum_{j \neq i} a_{ij} x_j^r + b_i.$$  

(15.8)

In terms of matrices, letting

$$A = \begin{pmatrix} * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{pmatrix}$$

The iterates are defined as

$$\begin{pmatrix} * & 0 & \cdots & 0 \\ 0 & * & \ddots & \vdots \\ \vdots & \ddots & * & 0 \\ 0 & \cdots & 0 & * \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ \vdots \\ x_n^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & * & \cdots & * \\ \vdots & 0 & \ddots & \vdots \\ * & \ddots & * & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ \vdots \\ x_n^r \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$  

(15.9)

The matrix on the left in (15.9) is obtained by retaining the main diagonal of $A$ and setting every other entry equal to zero. The matrix on the right in (15.9) is obtained from $A$ by setting every diagonal entry equal to zero and retaining all the other entries unchanged.

**Example 15.4.2** Use the Jacobi method to solve the system

$$\begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 2 & 5 & 1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$
CHAPTER 15. ANALYSIS OF LINEAR TRANSFORMATIONS

Of course this is solved most easily using row reductions. The Jacobi method is useful when the matrix is very large. This example is just to illustrate how the method works. First let's solve it using row operations. The exact solution from row reduction is $egin{pmatrix} \frac{6}{29} & \frac{11}{29} & \frac{8}{29} & \frac{25}{29} \end{pmatrix}$, which in terms of decimals is approximately equal to

$$
\begin{pmatrix}
0.207 & 0.379 & 0.276 & 0.862
\end{pmatrix}^T.
$$

In terms of the matrices, the Jacobi iteration is of the form

$$
\begin{pmatrix}
x_1^{r+1} \\
x_2^{r+1} \\
x_3^{r+1} \\
x_4^{r+1}
\end{pmatrix} = - \begin{pmatrix}
0 & 1/3 & 0 & 0 \\
1/4 & 0 & 1/4 & 0 \\
0 & 1/2 & 0 \\
0 & 0 & 1/2 & 0
\end{pmatrix} \begin{pmatrix}
x_1^r \\
x_2^r \\
x_3^r \\
x_4^r
\end{pmatrix} + \begin{pmatrix}
1/3 \\
1/4 \\
0 \\
0
\end{pmatrix}.
$$

(15.10)

Multiplying by the inverse of the matrix on the left, this iteration reduces to

$$
\begin{pmatrix}
x_1^{r+1} \\
x_2^{r+1} \\
x_3^{r+1} \\
x_4^{r+1}
\end{pmatrix} = \begin{pmatrix}
0 & 1/3 & 0 & 0 \\
1/4 & 0 & 1/4 & 0 \\
0 & 1/2 & 0 \\
0 & 0 & 1/2 & 0
\end{pmatrix} \begin{pmatrix}
x_2^r \\
x_3^r \\
x_4^r + 1
\end{pmatrix}.
$$

Now iterate this starting with $x^1 \equiv \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}^T$.

Thus

$$
x^2 = - \begin{pmatrix}
0 & 1/3 & 0 & 0 \\
1/4 & 0 & 1/4 & 0 \\
0 & 1/2 & 0 \\
0 & 0 & 1/2 & 0
\end{pmatrix} \begin{pmatrix}
x_2^1 \\
x_3^1 \\
x_4^1 \\
x_5^1
\end{pmatrix} + \begin{pmatrix}
1/3 \\
1/4 \\
0 \\
0
\end{pmatrix} = \begin{pmatrix}
0.166 \\
0.26 \\
0.2 \\
0.7
\end{pmatrix}.
$$

Then

$$
x^3 = - \begin{pmatrix}
0 & 1/3 & 0 & 0 \\
1/4 & 0 & 1/4 & 0 \\
0 & 1/2 & 0 \\
0 & 0 & 1/2 & 0
\end{pmatrix} \begin{pmatrix}
x_2^2 \\
x_3^2 \\
x_4^2 + 1
\end{pmatrix} + \begin{pmatrix}
1/3 \\
1/4 \\
0 \\
0
\end{pmatrix} = \begin{pmatrix}
.197 \\
.351 \\
.2566 \\
.822
\end{pmatrix}.
$$

Continuing this way one finally gets

$$
x^6 = - \begin{pmatrix}
0 & 1/3 & 0 & 0 \\
1/4 & 0 & 1/4 & 0 \\
0 & 1/2 & 0 \\
0 & 0 & 1/2 & 0
\end{pmatrix} \begin{pmatrix}
x_2^5 \\
x_3^5 \\
x_4^5 + 1
\end{pmatrix} + \begin{pmatrix}
1/3 \\
1/4 \\
0 \\
0
\end{pmatrix} = \begin{pmatrix}
.216 \\
.386 \\
.295 \\
.871
\end{pmatrix}.
$$

You can keep going like this. Recall the solution is approximately equal to

$$
\begin{pmatrix}
0.207 & 0.379 & 0.276 & 0.862
\end{pmatrix}^T.
$$

so you see that with no care at all and only 6 iterations, an approximate solution has been obtained which is not too far off from the actual solution.

\footnote{You certainly would not compute the inverse in solving a large system. This is just to show you how the method works for this simple example. You would use the first description in terms of indices.}


**Definition 15.4.3** The Gauss Seidel method, also called the method of successive corrections is given as follows. For $A = (a_{ij})$, the iterates for the problem $Ax = b$ are obtained according to the formula

$$\sum_{j=1}^{i} a_{ij} x_{j}^{r+1} = -\sum_{j=i+1}^{n} a_{ij} x_{j}^{r} + b_{i}. \quad (15.11)$$

In terms of matrices, letting

$$A = \begin{pmatrix} * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{pmatrix},$$

the iterates are defined as

$$\begin{pmatrix} * & 0 & \cdots & 0 \\ * & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \cdots & * & * \end{pmatrix} \begin{pmatrix} x_{1}^{r+1} \\ x_{2}^{r+1} \\ \vdots \\ x_{n}^{r+1} \end{pmatrix} = -\begin{pmatrix} 0 & * & \cdots & * \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & 0 \end{pmatrix} \begin{pmatrix} x_{1}^{r} \\ x_{2}^{r} \\ x_{3}^{r} \\ x_{n}^{r} \end{pmatrix} + \begin{pmatrix} b_{1} \\ b_{2} \\ \vdots \\ b_{n} \end{pmatrix} \quad (15.12)$$

In words, you set every entry in the original matrix which is strictly above the main diagonal equal to zero to obtain the matrix on the left. To get the matrix on the right, you set every entry of $A$ which is on or below the main diagonal equal to zero. Using the iteration procedure directly, the Gauss Seidel method makes use of the very latest information which is available at that stage of the computation.

The following example is the same as the example used to illustrate the Jacobi method.

**Example 15.4.4** Use the Gauss Seidel method to solve the system

$$\begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 2 & 5 & 1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_{1} \\ x_{2} \\ x_{3} \\ x_{4} \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

In terms of matrices, this procedure is

$$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 0 & 2 & 5 & 0 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_{1}^{r+1} \\ x_{2}^{r+1} \\ x_{3}^{r+1} \\ x_{4}^{r+1} \end{pmatrix} = -\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_{1}^{r} \\ x_{2}^{r} \\ x_{3}^{r} \\ x_{n}^{r} \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

Multiplying by the inverse of the matrix on the left\(^2\) this yields

$$\begin{pmatrix} x_{1}^{r+1} \\ x_{2}^{r+1} \\ x_{3}^{r+1} \\ x_{4}^{r+1} \end{pmatrix} = -\begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{12} & \frac{1}{4} & 0 \\ 0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} x_{1}^{r} \\ x_{2}^{r} \\ x_{3}^{r} \\ x_{4}^{r} \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix}.$$

\(^2\)As in the case of the Jacobi iteration, the computer would not do this. It would use the iteration procedure in terms of the entries of the matrix directly. Otherwise all benefit to using this method is lost.
As before, I will be totally unoriginal in the choice of $x^1$. Let it equal the zero vector. Therefore, $x^2 = \left( \frac{1}{3} \; \frac{5}{12} \; \frac{13}{30} \; \frac{47}{60} \right)^T$. Now,

$$x^3 = -\begin{pmatrix}
 0 & \frac{1}{3} & 0 & 0 \\
 0 & -\frac{1}{12} & \frac{1}{4} & 0 \\
 0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\
 0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10}
\end{pmatrix} \begin{pmatrix}
 0.194 \\
 0.343 \\
 0.306 \\
 0.846
\end{pmatrix} \begin{pmatrix}
 0.194 \\
 0.343 \\
 0.306 \\
 0.846
\end{pmatrix} + \begin{pmatrix}
 0.194 \\
 0.343 \\
 0.306 \\
 0.846
\end{pmatrix} = \begin{pmatrix}
 0.194 \\
 0.343 \\
 0.306 \\
 0.846
\end{pmatrix}.$$

Continuing this way,

$$x^4 = -\begin{pmatrix}
 0 & \frac{1}{3} & 0 & 0 \\
 0 & -\frac{1}{12} & \frac{1}{4} & 0 \\
 0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\
 0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10}
\end{pmatrix} \begin{pmatrix}
 0.219 \\
 0.36875 \\
 0.283 \\
 0.85835
\end{pmatrix} \begin{pmatrix}
 0.219 \\
 0.36875 \\
 0.283 \\
 0.85835
\end{pmatrix} + \begin{pmatrix}
 0.219 \\
 0.36875 \\
 0.283 \\
 0.85835
\end{pmatrix} = \begin{pmatrix}
 0.219 \\
 0.36875 \\
 0.283 \\
 0.85835
\end{pmatrix}.$$

and so

$$x^5 = -\begin{pmatrix}
 0 & \frac{1}{3} & 0 & 0 \\
 0 & -\frac{1}{12} & \frac{1}{4} & 0 \\
 0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\
 0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10}
\end{pmatrix} \begin{pmatrix}
 0.219 \\
 0.36875 \\
 0.283 \\
 0.85835
\end{pmatrix} \begin{pmatrix}
 0.219 \\
 0.36875 \\
 0.283 \\
 0.85835
\end{pmatrix} + \begin{pmatrix}
 0.219 \\
 0.36875 \\
 0.283 \\
 0.85835
\end{pmatrix} = \begin{pmatrix}
 0.21042 \\
 0.37657 \\
 0.2777 \\
 0.86115
\end{pmatrix}.$$

This is fairly close to the answer. You could continue doing these iterates and it appears they converge to the solution. Now consider the following example.

**Example 15.4.5** Use the Gauss Seidel method to solve the system

$$\begin{pmatrix}
 1 & 4 & 0 & 0 \\
 1 & 4 & 1 & 0 \\
 0 & 2 & 5 & 1 \\
 0 & 0 & 2 & 4
\end{pmatrix} \begin{pmatrix}
 x_1 \\
 x_2 \\
 x_3 \\
 x_4
\end{pmatrix} = \begin{pmatrix}
 1 \\
 2 \\
 3 \\
 4
\end{pmatrix}.$$

The exact solution is given by doing row operations on the augmented matrix. When this is done the solution is seen to be $\left( \frac{6.0}{6.0} \; -1.25 \; 1.0 \; 0.5 \right)$.The Gauss Seidel iterations are of the form

$$\begin{pmatrix}
 1 & 0 & 0 & 0 \\
 1 & 4 & 0 & 0 \\
 0 & 2 & 5 & 0 \\
 0 & 0 & 2 & 4
\end{pmatrix} \begin{pmatrix}
 x_1^{r+1} \\
 x_2^{r+1} \\
 x_3^{r+1} \\
 x_4^{r+1}
\end{pmatrix} = -\begin{pmatrix}
 0 & 4 & 0 & 0 \\
 0 & 0 & 1 & 0 \\
 0 & 0 & 0 & 1 \\
 0 & 0 & 0 & 0
\end{pmatrix} \begin{pmatrix}
 x_1^r \\
 x_2^r \\
 x_3^r \\
 x_4^r
\end{pmatrix} + \begin{pmatrix}
 1 \\
 2 \\
 3 \\
 4
\end{pmatrix}$$

and so, multiplying by the inverse of the matrix on the left, the iteration reduces to the following in terms of matrix multiplication.

$$x^{r+1} = -\begin{pmatrix}
 0 & 4 & 0 & 0 \\
 0 & -1 & \frac{1}{4} & 0 \\
 0 & 2 & \frac{1}{5} & -\frac{1}{10} \\
 0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10}
\end{pmatrix} x^r + \begin{pmatrix}
 1 \\
 \frac{1}{2} \\
 \frac{1}{2} \\
 \frac{1}{2}
\end{pmatrix}.$$
This time, I will pick an initial vector close to the answer. Let \( \mathbf{x}^1 = \begin{pmatrix} 6 & -1 & 1 & \frac{1}{2} \end{pmatrix}^T \). This is very close to the answer. Now lets see what the Gauss Seidel iteration does to it.

\[
x^2 = - \begin{pmatrix} 0 & 4 & 0 & 0 \\ 0 & -1 & \frac{4}{10} & 0 \\ 0 & \frac{2}{5} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} 6 \\ -1 \\ 1 \\ \frac{1}{2} \end{pmatrix} + \begin{pmatrix} 1 \\ \frac{1}{4} \\ \frac{1}{2} \\ \frac{3}{4} \end{pmatrix} = \begin{pmatrix} 5.0 \\ -1.0 \\ .9 \\ .55 \end{pmatrix}
\]

It appears that it moved the initial guess far from the solution even though you started with one which was initially close to the solution. This is discouraging. However, you can’t expect the method to work well after only one iteration. Unfortunately, if you do multiple iterations, the iterates never seem to get close to the actual solution. Why is the process which worked so well in the other examples not working here? A better question might be: Why does either process ever work at all?

Both iterative procedures for solving \( A \mathbf{x} = \mathbf{b} \)

are of the form

\[ B \mathbf{x}^{r+1} = -C \mathbf{x}^r + \mathbf{b} \]

where \( A = B + C \). In the Jacobi procedure, the matrix \( C \) was obtained by setting the diagonal of \( A \) equal to zero and leaving all other entries the same while the matrix \( B \) was obtained by making every entry of \( A \) equal to zero other than the diagonal entries which are left unchanged. In the Gauss Seidel procedure, the matrix \( B \) was obtained from \( A \) by making every entry strictly above the main diagonal equal to zero and leaving the others unchanged, and \( C \) was obtained from \( A \) by making every entry on or below the main diagonal equal to zero and leaving the others unchanged. Thus in the Jacobi procedure, \( B \) is a diagonal matrix while in the Gauss Seidel procedure, \( B \) is lower triangular. Using matrices to explicitly solve for the iterates, yields

\[
\mathbf{x}^{r+1} = -B^{-1}C\mathbf{x}^r + B^{-1}\mathbf{b}.
\quad (15.14)
\]

This is what you would never have the computer do but this is what will allow the statement of a theorem which gives the condition for convergence of these and all other similar methods. Recall the definition of the spectral radius of \( M, \rho (M) \), in Definition \( 15.2.1 \) on Page 354.

**Theorem 15.4.6** Suppose \( \rho (B^{-1}C) < 1 \). Then the iterates in (15.14) converge to the unique solution of (15.13).

I will prove this theorem in the next section. The proof depends on analysis which should not be surprising because it involves a statement about convergence of sequences.

What is an easy to verify sufficient condition which will imply the above holds? It is easy to give one in the case of the Jacobi method. Suppose the matrix \( A \) is diagonally dominant. That is \( |a_{ii}| > \sum_{j \neq i} |a_{ij}| \). Then \( B \) would be the diagonal matrix consisting of the entries \( a_{ii} \). You need to find the size of \( \lambda \) where

\[
B^{-1}C\mathbf{x} = \lambda \mathbf{x}
\]

Thus you need

\[
(\lambda B - C)\mathbf{x} = \mathbf{0}
\]

Now if \( |\lambda| \geq 1 \), then the matrix \( \lambda B - C \) is diagonally dominant and so this matrix will be invertible so \( \lambda \) is not an eigenvalue. Hence the only eigenvalues have absolute value less than 1.

You might try a similar argument in the case of the Gauss Seidel method.
15.5 Theory Of Convergence

Definition 15.5.1 A normed vector space, $E$ with norm $||\cdot||$ is called a Banach space if it is also complete. This means that every Cauchy sequence converges. Recall that a sequence $\{x_n\}_{n=1}^{\infty}$ is a Cauchy sequence if for every $\varepsilon > 0$ there exists $N$ such that whenever $m, n > N$,

$$||x_n - x_m|| < \varepsilon.$$  

Thus whenever $\{x_n\}$ is a Cauchy sequence, there exists $x$ such that

$$\lim_{n \to \infty} ||x - x_n|| = 0.$$  

Example 15.5.2 Let $E$ be a Banach space and let $\Omega$ be a nonempty subset of a normed linear space $F$. Let $B(\Omega; E)$ denote those functions $f$ for which

$$||f|| = \sup \{||f(x)||_E : x \in \Omega\} < \infty$$

Denote by $BC(\Omega; E)$ the set of functions in $B(\Omega; E)$ which are also continuous.

Lemma 15.5.3 The above $||\cdot||$ is a norm on $B(\Omega; E)$. The subspace $BC(\Omega; E)$ with the given norm is a Banach space.

Proof: It is obvious $||\cdot||$ is a norm. It only remains to verify $BC(\Omega; E)$ is complete. Let $\{f_n\}$ be a Cauchy sequence. Since $||f_n - f_m|| \to 0$ as $m, n \to \infty$, it follows that $\{f_n(x)\}$ is a Cauchy sequence in $E$ for each $x$. Let $f(x) = \lim_{n \to \infty} f_n(x)$. Then for any $x \in \Omega$,

$$||f_n(x) - f_m(x)||_E \leq ||f_n - f_m|| < \varepsilon$$

whenever $m, n$ are large enough, say as large as $N$. For $n \geq N$, let $m \to \infty$. Then passing to the limit, it follows that for all $x$,

$$||f_n(x) - f(x)||_E \leq \varepsilon$$

and so for all $x$,

$$||f(x)||_E \leq \varepsilon + ||f_n(x)||_E \leq \varepsilon + ||f_n||.$$

It follows that $||f|| \leq ||f_n|| + \varepsilon$ and $||f - f_n|| \leq \varepsilon$.

It remains to verify that $f$ is continuous.

\[
\begin{align*}
||f(x) - f(y)||_E & \leq ||f(x) - f_n(x)||_E + ||f_n(x) - f_n(y)||_E + ||f_n(y) - f(y)||_E \\
& \leq 2||f - f_n|| + ||f_n(x) - f_n(y)||_E \leq \frac{2\varepsilon}{3} + ||f_n(x) - f_n(y)||_E
\end{align*}
\]

for all $n$ large enough. Now pick such an $n$. By continuity, the last term is less than $\frac{\varepsilon}{3}$ if $||x - y||$ is small enough. Hence $f$ is continuous as well.

The most familiar example of a Banach space is $\mathbb{R}^n$. The following lemma is of great importance so it is stated in general.

Lemma 15.5.4 Suppose $T : E \to E$ where $E$ is a Banach space with norm $||\cdot||$. Also suppose

$$||Tx - Ty|| \leq r||x - y||$$  

for some $r \in (0, 1)$. Then there exists a unique fixed point, $x \in E$ such that

$$Tx = x.$$  

Letting $x^1 \in E$, this fixed point $x$, is the limit of the sequence of iterates,

$$x^1, Tx^1, T^2x^1, \ldots.$$  

In addition to this, there is a nice estimate which tells how close $x^1$ is to $x$ in terms of things which can be computed.

$$||x^1 - x|| \leq \frac{1}{1 - r} ||x^1 - Tx^1||.$$  

(15.18)
Proof: This follows easily when it is shown that the above sequence, \( \{T^k x^1\}_{k=1}^{\infty} \) is a Cauchy sequence. Note that
\[
|T^2 x^1 - T x^1| \leq r |T x^1 - x^1|.
\]
Suppose
\[
|T^k x^1 - T^{k-1} x^1| \leq r^{k-1} |T x^1 - x^1|.
\] (15.19)
Then
\[
|T^{k+1} x^1 - T^k x^1| \leq r |T^k x^1 - T^{k-1} x^1| \\
\leq r r^{k-1} |T x^1 - x^1| = r^k |T x^1 - x^1|.
\]
By induction, this shows that for all \( k \geq 2 \), (15.19) is valid. Now let \( k > l \geq N \).
\[
|T^k x^1 - T^l x^1| = \left| \sum_{j=l}^{k-1} (T^{j+1} x^1 - T^j x^1) \right| \leq \sum_{j=l}^{k-1} |T^{j+1} x^1 - T^j x^1| \\
\leq \sum_{j=N}^{k-1} r^j |T x^1 - x^1| \leq |T x^1 - x^1| \frac{r^N}{1-r}
\]
which converges to 0 as \( N \to \infty \). Therefore, this is a Cauchy sequence so it must converge to \( x \in E \).
Then
\[
x = \lim_{k \to \infty} T^k x^1 = \lim_{k \to \infty} T^{k+1} x^1 = T \lim_{k \to \infty} T^k x^1 = Tx.
\]
This shows the existence of the fixed point. To show it is unique, suppose there were another one, \( y \). Then
\[
|x - y| = |T x - T y| \leq r |x - y|
\]
and so \( x = y \).

It remains to verify the estimate.
\[
|x^1 - x| \leq |x^1 - T x^1| + |T x^1 - x| = |x^1 - T x^1| + |T x^1 - T x|
\]
\[
\leq |x^1 - T x^1| + r |x^1 - x|
\]
and solving the inequality for \( |x^1 - x| \) gives the estimate desired. □

The following corollary is what will be used to prove the convergence condition for the various iterative procedures.

Corollary 15.5.5 Suppose \( T : E \to E \), for some constant \( C \)
\[
|T x - T y| \leq C |x - y|,
\]
for all \( x, y \in E \), and for some \( N \in \mathbb{N} \),
\[
|T^N x - T^N y| \leq r |x - y|,
\]
for all \( x, y \in E \) where \( r \in (0, 1) \). Then there exists a unique fixed point for \( T \) and it is still the limit of the sequence, \( \{ T^k x^1 \} \) for any choice of \( x^1 \).

Proof: From Lemma 15.5.4 there exists a unique fixed point for \( T^N \) denoted here as \( x \). Therefore, \( T^N x = x \). Now doing \( T \) to both sides,
\[
T^N T x = T x.
\]
By uniqueness, \( T x = x \) because the above equation shows \( T x \) is a fixed point of \( T^N \) and there is only one fixed point of \( T^N \). In fact, there is only one fixed point of \( T \) because a fixed point of \( T \) is automatically a fixed point of \( T^N \).
It remains to show $T^k x^1 \rightarrow x$, the unique fixed point of $T^N$. If this does not happen, there exists $\varepsilon > 0$ and a subsequence, still denoted by $T^k$ such that

$$|T^k x^1 - x| \geq \varepsilon$$

Now $k = j_k N + r_k$ where $r_k \in \{0, \cdots , N - 1\}$ and $j_k$ is a positive integer such that $\lim_{k \to \infty} j_k = \infty$. Then there exists a single $r \in \{0, \cdots , N - 1\}$ such that for infinitely many $k, r_k = r$. Taking a further subsequence, still denoted by $T^k$ it follows

$$|T^{j_k N + r} x^1 - x| \geq \varepsilon \quad (15.20)$$

However,

$$T^{j_k N + r} x^1 = T^r T^{j_k N} x^1 \rightarrow T^r x = x$$

and this contradicts (15.20).

**Theorem 15.5.6** Suppose $\rho (B^{-1}C) < 1$. Then the iterates in (15.14) converge to the unique solution of (15.13).

**Proof:** Consider the iterates in (15.14). Let $T x = B^{-1} C x + B^{-1} b$. Then

$$|T^k x - T^k y| = \left| (B^{-1} C)^k x - (B^{-1} C)^k y \right| \leq \left| |(B^{-1} C)^k|| |x - y| \right| .$$

Here $|| \cdot ||$ refers to any of the operator norms. It doesn’t matter which one you pick because they are all equivalent. I am writing the proof to indicate the operator norm taken with respect to the usual norm on $E$. Since $\rho (B^{-1}C) < 1$, it follows from Gelfand’s theorem, Theorem 15.2.3 on Page 356, there exists $N$ such that if $k \geq N$, then for some $r^{1/k} < 1$,

$$\left| |(B^{-1} C)^k|| \right|^{1/k} < r^{1/k} < 1.$$

Consequently,

$$|T^N x - T^N y| \leq r |x - y| .$$

Also $|T x - T y| \leq ||B^{-1}C|| |x - y|$ and so Corollary 15.5.2 applies and gives the conclusion of this theorem. ■

### 15.6 Exercises

1. Solve the system

$$\begin{pmatrix} 4 & 1 & 1 \\ 1 & 5 & 2 \\ 0 & 2 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

2. Solve the system

$$\begin{pmatrix} 4 & 1 & 1 \\ 1 & 7 & 2 \\ 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.
3. Solve the system
\[
\begin{pmatrix}
5 & 1 & 1 \\
1 & 7 & 2 \\
0 & 2 & 4
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= 
\begin{pmatrix}
1 \\
2 \\
3
\end{pmatrix}
\]
using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

4. If you are considering a system of the form \( Ax = b \) and \( A^{-1} \) does not exist, will either the Gauss Seidel or Jacobi methods work? Explain. What does this indicate about finding eigenvectors for a given eigenvalue?

5. For \( \| x \|_\infty \equiv \max \{ |x_j| : j = 1, 2, \cdots, n \} \), the parallelogram identity does not hold. Explain.

6. A norm \( \| \cdot \| \) is said to be strictly convex if whenever \( \| x \| = \| y \|, x \neq y \), it follows
\[
\frac{x + y}{2} < \| x \| = \| y \|.
\]
Show the norm \( \| \cdot \| \) which comes from an inner product is strictly convex.

7. A norm \( \| \cdot \| \) is said to be uniformly convex if whenever \( \| x_n \|, \| y_n \| \) are equal to 1 for all \( n \in \mathbb{N} \) and \( \lim_{n \to \infty} \| x_n + y_n \| = 2 \), it follows \( \lim_{n \to \infty} \| x_n - y_n \| = 0 \). Show the norm \( \| \cdot \| \) coming from an inner product is always uniformly convex. Also show that uniform convexity implies strict convexity which is defined in Problem 6.

8. Suppose \( A : \mathbb{C}^n \to \mathbb{C}^n \) is a one to one and onto matrix. Define
\[
\| x \| \equiv |Ax|.
\]
Show this is a norm.

9. If \( X \) is a finite dimensional normed vector space and \( A, B \in \mathcal{L}(X, X) \) such that \( \| B \| < \| A \| \), can it be concluded that \( \| A^{-1} B \| < 1 \)?

10. Let \( X \) be a vector space with a norm \( \| \cdot \| \) and let \( V = \text{span}(v_1, \cdots, v_m) \) be a finite dimensional subspace of \( X \) such that \( \{v_1, \cdots, v_m\} \) is a basis for \( V \). Show \( V \) is a closed subspace of \( X \). This means that if \( w_n \to w \) and each \( w_n \in V \), then so is \( w \). Next show that if \( w \notin V \),
\[
\text{dist}(w, V) \equiv \inf \{ \| w - v \| : v \in V \} > 0
\]
is a continuous function of \( w \) and
\[
|\text{dist}(w, V) - \text{dist}(w_1, V)| \leq \| w_1 - w \|
\]
Next show that if \( w \notin V \), there exists \( z \) such that \( \| z \| = 1 \) and \( \text{dist}(z, V) > 1/2 \). For those who know some advanced calculus, show that if \( X \) is an infinite dimensional vector space having norm \( \| \cdot \| \), then the closed unit ball in \( X \) cannot be compact. Thus closed and bounded is never compact in an infinite dimensional normed vector space.

11. Suppose \( \rho(A) < 1 \) for \( A \in \mathcal{L}(V, V) \) where \( V \) is a \( p \) dimensional vector space having a norm \( \| \cdot \| \). You can use \( \mathbb{R}^p \) or \( \mathbb{C}^p \) if you like. Show there exists a new norm \( \| \| \cdot \| \| \) such that with respect to this new norm, \( \| \| A \| \| < 1 \) where \( \| A \| \) denotes the operator norm of \( A \) taken with respect to this new norm on \( V \),
\[
\| A \| \equiv \sup \{ \| A x \| : \| x \| \leq 1 \}
\]
\textbf{Hint:} You know from Gelfand’s theorem that
\[
\| A^n \|^{1/n} < r < 1
\]
provided \( n \) is large enough, this operator norm taken with respect to \( \| \cdot \| \). Show there exists
\( 0 < \lambda < 1 \) such that
\[
\rho \left( \frac{A}{\lambda} \right) < 1.
\]
You can do this by arguing the eigenvalues of \( A/\lambda \) are the scalars \( \mu/\lambda \) where \( \mu \in \sigma (A) \). Now let \( \mathbb{Z}_+ \) denote the nonnegative integers.

\[
\|\| x \|\| \equiv \sup_{n \in \mathbb{Z}_+} \left\| \frac{A^n}{\lambda^n} x \right\|
\]
First show this is actually a norm. Next explain why
\[
\|A x\| \equiv \lambda \sup_{n \in \mathbb{Z}_+} \left| \frac{A^{n+1}}{\lambda^{n+1}} x \right| \leq \lambda \|x\|.
\]

12. Establish a similar result to Problem 11 without using Gelfand’s theorem. Use an argument which depends directly on the Jordan form or a modification of it.

13. Using Problem 11 give an easier proof of Theorem 15.5.1 without having to use Corollary 15.5.5. It would suffice to use a different norm of this problem and the contraction mapping principle of Lemma 15.5.4.

14. A matrix \( A \) is diagonally dominant if \( |a_{ii}| > \sum_{j \neq i} |a_{ij}| \). Show that the Gauss Seidel method converges if \( A \) is diagonally dominant.

15. Suppose \( f (\lambda) = \sum_{n=0}^{\infty} a_n \lambda^n \) converges if \( |\lambda| < R \). Show that if \( \rho (A) < R \) where \( A \) is an \( n \times n \) matrix, then
\[
f (A) = \sum_{n=0}^{\infty} a_n A^n
\]
converges in \( \mathcal{L}(\mathbb{F}^n, \mathbb{F}^n) \). Hint: Use Gelfand’s theorem and the root test.

16. Referring to Corollary 15.3.5 for \( \lambda = a + ib \) show
\[
\exp (\lambda t) = e^{at} (\cos (bt) + i \sin (bt)).
\]

Hint: Let \( y (t) = \exp (\lambda t) \) and let \( z (t) = e^{-at} y (t) \). Show
\[
z'' + b^2 z = 0, \quad z (0) = 1, \quad z' (0) = ib.
\]
Now letting \( z = u + iv \) where \( u, v \) are real valued, show
\[
\begin{align*}
u'' + b^2 u &= 0, \quad u (0) = 1, \quad u' (0) = 0 \\
v'' + b^2 v &= 0, \quad v (0) = 0, \quad v' (0) = b.
\end{align*}
\]
Next show \( u (t) = \cos (bt) \) and \( v (t) = \sin (bt) \) work in the above and that there is at most one solution to
\[
w'' + b^2 w = 0 \quad w (0) = \alpha, \quad w' (0) = \beta.
\]
Thus \( z (t) = \cos (bt) + i \sin (bt) \) and so \( y (t) = e^{at} (\cos (bt) + i \sin (bt)) \). To show there is at most one solution to the above problem, suppose you have two, \( w_1, w_2 \). Subtract them. Let \( f = w_1 - w_2 \). Thus
\[
f'' + b^2 f = 0
\]
and \( f \) is real valued. Multiply both sides by \( f' \) and conclude
\[
\frac{d}{dt} \left( \frac{(f')^2}{2} + \frac{b^2 f^2}{2} \right) = 0
\]
Thus the expression in parenthesis is constant. Explain why this constant must equal 0.
17. Let \( A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) \). Show the following power series converges in \( \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) \).

\[
\Psi(t) = \sum_{k=0}^{\infty} \frac{t^k A^k}{k!}
\]

This was done in the chapter. Go over it and be sure you understand it. This is how you can define \( \exp(tA) \). Next show that \( \Psi'(t) = A\Psi(t) \), \( \Psi(0) = I \). Next let \( \Phi(t) = \sum_{k=0}^{\infty} \frac{t^k (-A)^k}{k!} \). Show each \( \Phi(t), \Psi(t) \) each commute with \( A \). Next show that \( \Phi(t) \Psi(t) = I \) for all \( t \). Finally, solve the initial value problem

\[
x' = Ax + f, \quad x(0) = x_0
\]

in terms of \( \Phi \) and \( \Psi \). This yields most of the substance of a typical differential equations course.

18. In Problem \( \square \) \( \Psi(t) \) is defined by the given series. Denote by \( \exp(t\sigma(A)) \) the numbers \( \exp(t\lambda) \) where \( \lambda \in \sigma(A) \). Show \( \exp(t\sigma(A)) = \sigma(\Psi(t)) \). This is like Lemma \( \square \). Letting \( J \) be the Jordan canonical form for \( A \), explain why

\[
\Psi(t) = \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} = S \sum_{k=0}^{\infty} \frac{t^k J^k}{k!} S^{-1}
\]

and you note that in \( J^k \), the diagonal entries are of the form \( \lambda^k \) for \( \lambda \) an eigenvalue of \( A \). Also \( J = D + N \) where \( N \) is nilpotent and commutes with \( D \). Argue then that

\[
\sum_{k=0}^{\infty} \frac{t^k J^k}{k!}
\]

is an upper triangular matrix which has on the diagonal the expressions \( e^{\lambda t} \) where \( \lambda \in \sigma(A) \). Thus conclude

\[
\sigma(\Psi(t)) \subseteq \exp(t\sigma(A))
\]

Next take \( e^{t\lambda} \in \exp(t\sigma(A)) \) and argue it must be in \( \sigma(\Psi(t)) \). You can do this as follows:

\[
\Psi(t) - e^{t\lambda}I = \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} - \sum_{k=0}^{\infty} \frac{t^k \lambda^k}{k!} I = \sum_{k=0}^{\infty} \frac{t^k}{k!} \left( A^k - \lambda^k I \right)
\]

\[
= \left( \sum_{k=0}^{\infty} \frac{t^k}{k!} \sum_{j=1}^{k-1} A^{k-j} \lambda^j \right) (A - \lambda I)
\]

Now you need to argue

\[
\sum_{k=0}^{\infty} \frac{t^k}{k!} \sum_{j=1}^{k-1} A^{k-j} \lambda^j
\]

converges to something in \( \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) \). To do this, use the ratio test and Lemma \( \square \) after first using the triangle inequality. Since \( \lambda \in \sigma(A) \), \( \Psi(t) - e^{t\lambda}I \) is not one to one and so this establishes the other inclusion. You fill in the details. This theorem is a special case of theorems which go by the name “spectral mapping theorem” which was discussed in the text. However, go through it yourself.

19. Suppose \( \Psi(t) \in \mathcal{L}(V, W) \) where \( V, W \) are finite dimensional inner product spaces and \( t \to \Psi(t) \) is continuous for \( t \in [a, b] \). For every \( \varepsilon > 0 \) there there exists \( \delta > 0 \) such that if \( |s - t| < \delta \) then \( ||\Psi(t) - \Psi(s)|| < \varepsilon \). Show \( t \to (\Psi(t)v, w) \) is continuous. Here it is the inner product in \( W \). Also define what it means for \( t \to \Psi(t)v \) to be continuous and show this is continuous. Do it all for differentiable in place of continuous. Next show \( t \to ||\Psi(t)|| \) is continuous.
20. If \( z(t) \in W \), a finite dimensional inner product space, what does it mean for \( t \to z(t) \) to be continuous or differentiable? If \( z \) is continuous, define
\[
\int_a^b z(t) \, dt \in W
\]
as follows.
\[
\left( w, \int_a^b z(t) \, dt \right) = \int_a^b (w, z(t)) \, dt.
\]
Show that this definition is well defined and furthermore the triangle inequality,
\[
\left\| \int_a^b z(t) \, dt \right\| \leq \int_a^b \| z(t) \| \, dt,
\]
and fundamental theorem of calculus,
\[
\frac{d}{dt} \left( \int_a^t z(s) \, ds \right) = z(t)
\]
hold along with any other interesting properties of integrals which are true.

21. For \( V, W \) two inner product spaces, define
\[
\int_a^b \Psi(t) \, dt \in \mathcal{L}(V, W)
\]
as follows.
\[
\left( w, \int_a^b \Psi(t) \, dt \, (v) \right) = \int_a^b (w, \Psi(t) \, v) \, dt.
\]
Show this is well defined and does indeed give \( \int_a^b \Psi(t) \, dt \in \mathcal{L}(V, W) \). Also show the triangle inequality
\[
\left\| \int_a^b \Psi(t) \, dt \right\| \leq \int_a^b \| \Psi(t) \| \, dt
\]
where \( \| \cdot \| \) is the operator norm and verify the fundamental theorem of calculus holds.
\[
\left( \int_a^t \Psi(s) \, ds \right)' = \Psi(t).
\]
Also verify the usual properties of integrals continue to hold such as the fact the integral is linear and
\[
\int_a^b \Psi(t) \, dt + \int_b^c \Psi(t) \, dt = \int_a^c \Psi(t) \, dt
\]
and similar things. **Hint:** On showing the triangle inequality, it will help if you use the fact that
\[
|w|_W = \sup_{\|v\| \leq 1} |(w, v)|.
\]
You should show this also.

22. Prove Gronwall’s inequality. Suppose \( u(t) \geq 0 \) and for all \( t \in [0, T] \),
\[
u(t) \leq u_0 + \int_0^t K u(s) \, ds.
\]
where $K$ is some nonnegative constant. Then

$$u(t) \leq u_0 e^{Kt}.$$  

**Hint:** $w(t) = \int_0^t u(s) \, ds$. Then using the fundamental theorem of calculus, $w(t)$ satisfies the following.

$$u(t) - Kw(t) = w'(t) - Kw(t) \leq u_0, \quad w(0) = 0.$$  

Now use the usual techniques you saw in an introductory differential equations class. Multiply both sides of the above inequality by $e^{-Kt}$ and note the resulting left side is now a total derivative. Integrate both sides from 0 to $t$ and see what you have got.

23. With Gronwall’s inequality and the integral defined in Problem 21 with its properties listed there, prove there is at most one solution to the initial value problem

$$y' = Ay, \quad y(0) = y_0.$$  

**Hint:** If there are two solutions, subtract them and call the result $z$. Then

$$z' = Az, \quad z(0) = 0.$$  

It follows

$$z(t) = 0 + \int_0^t A z(s) \, ds$$  

and so

$$||z(t)|| \leq \int_0^t ||A|| ||z(s)|| \, ds$$  

Now consider Gronwall’s inequality of Problem 22.

24. Suppose $A$ is a matrix which has the property that whenever $\mu \in \sigma(A)$, $\text{Re} \mu < 0$. Consider the initial value problem

$$y' = Ay, \quad y(0) = y_0.$$  

The existence and uniqueness of a solution to this equation has been established above in preceding problems, Problem 17 to 23. Show that in this case where the real parts of the eigenvalues are all negative, the solution to the initial value problem satisfies

$$\lim_{t \to \infty} y(t) = 0.$$  

**Hint:** A nice way to approach this problem is to show you can reduce it to the consideration of the initial value problem

$$z' = J_z z, \quad z(0) = z_0$$  

where $J_z$ is the modified Jordan canonical form where instead of ones down the main diagonal, there are $\varepsilon$ down the main diagonal (Problem 13). Then

$$z' = Dz + N_\varepsilon z$$  

where $D$ is the diagonal matrix obtained from the eigenvalues of $A$ and $N_\varepsilon$ is a nilpotent matrix commuting with $D$ which is very small provided $\varepsilon$ is chosen very small. Now let $\Psi(t)$ be the solution of

$$\Psi' = -D \Psi, \quad \Psi(0) = I$$  

described earlier as

$$\sum_{k=0}^{\infty} (-1)^k k! \frac{t^k D^k}{k!}.$$  

Thus $\Psi(t)$ commutes with $D$ and $N_\varepsilon$. Tell why. Next argue

$$(\Psi(t) z)' = \Psi(t) N_\varepsilon z(t)$$
and integrate from 0 to \( t \). Then

\[
\Psi (t) z(t) - z_0 = \int_0^t \Psi (s) N_z (s) \, ds.
\]

It follows

\[
||\Psi (t) z(t)|| \leq ||z_0|| + \int_0^t ||N_z|| ||\Psi (s) z(s)|| \, ds.
\]

It follows from Gronwall’s inequality

\[
||\Psi (t) z(t)|| \leq ||z_0|| e^{||N_z|| t}
\]

Now look closely at the form of \( \Psi (t) \) to get an estimate which is interesting. Explain why

\[
\Psi (t) = \begin{pmatrix} e^{\mu_1 t} & 0 \\ \vdots & \ddots \\ 0 & \cdots & e^{\mu_n t} \end{pmatrix}
\]

and now observe that if \( \varepsilon \) is chosen small enough, \( ||N_z|| \) is so small that each component of \( z(t) \) converges to 0.

25. Using Problem \ref{prob:24} show that if \( A \) is a matrix having the real parts of all eigenvalues less than 0 then if

\[
\Psi' (t) = A \Psi (t), \quad \Psi (0) = I
\]

it follows

\[
\lim_{t \to \infty} \Psi (t) = 0.
\]

**Hint:** Consider the columns of \( \Psi (t) \)?

26. Let \( \Psi (t) \) be a fundamental matrix satisfying

\[
\Psi' (t) = A \Psi (t), \quad \Psi (0) = I.
\]

Show \( \Psi (t)^n = \Psi (nt) \). **Hint:** Subtract and show the difference satisfies \( \Phi' = A \Phi, \quad \Phi (0) = 0 \). Use uniqueness.

27. If the real parts of the eigenvalues of \( A \) are all negative, show that for every positive \( t \),

\[
\lim_{n \to \infty} \Psi (nt) = 0.
\]

**Hint:** Pick \( \text{Re} (\sigma (A)) < -\lambda < 0 \) and use Problem \ref{prob:28} about the spectrum of \( \Psi (t) \) and Gelfand’s theorem for the spectral radius along with Problem \ref{prob:26} to argue that \( ||\Psi (nt) / e^{-\lambda nt}|| < 1 \) for all \( n \) large enough.

28. Let \( H \) be a Hermitian matrix. \( (H = H^*) \). Show that \( e^{iH} \equiv \sum_{n=0}^{\infty} \frac{(iH)^n}{n!} \) is unitary.

29. Show the converse of the above exercise. If \( V \) is unitary, then \( V = e^{iH} \) for some \( H \) Hermitian.

30. If \( U \) is unitary and does not have \(-1\) as an eigenvalue so that \( (I + U)^{-1} \) exists, show that

\[
H = i (I - U) (I + U)^{-1}
\]

is Hermitian. Then, verify that

\[
U = (I + iH) (I - iH)^{-1}.
\]

31. Suppose that \( A \in \mathcal{L} (V, V) \) where \( V \) is a normed linear space. Also suppose that \( ||A|| < 1 \) where this refers to the operator norm on \( A \). Verify that

\[
(I - A)^{-1} = \sum_{i=0}^{\infty} A^i.
\]

This is called the Neumann series. Suppose now that you only know the algebraic condition \( \rho (A) < 1 \). Is it still the case that the Neumann series converges to \( (I - A)^{-1} \)?
Chapter 16

Numerical Methods, Eigenvalues

16.1 The Power Method for Eigenvalues

This chapter discusses numerical methods for finding eigenvalues. However, to do this correctly, you must include numerical analysis considerations which are distinct from linear algebra. The purpose of this chapter is to give an introduction to some numerical methods without leaving the context of linear algebra. In addition, some examples are given which make use of computer algebra systems. For a more thorough discussion, you should see books on numerical methods in linear algebra like some listed in the references.

Let $A$ be a complex $p \times p$ matrix and suppose that it has distinct eigenvalues $\lambda_1, \ldots, \lambda_m$ and that $|\lambda_1| > |\lambda_k|$ for all $k$. Also let the Jordan form of $A$ be

$$J = \begin{pmatrix} J_1 & \cdots & \vdots \\ & \ddots & \vdots \\ & & J_m \end{pmatrix}$$

with $J_1$ an $m_1 \times m_1$ matrix.

$$J_k = \lambda_k I_k + N_k$$

where $N_k^r \neq 0$ but $N_k^{r+1} = 0$. Also let

$$P^{-1}AP = J, \; A = PJP^{-1}.$$ 

Now fix $x \in \mathbb{F}^p$. Take $Ax$ and let $s_1$ be the entry of the vector $Ax$ which has largest absolute value. Thus $Ax/s_1$ is a vector $y_1$ which has a component of 1 and every other entry of this vector has magnitude no larger than 1. If the scalars $\{s_1, \ldots, s_{n-1}\}$ and vectors $\{y_1, \ldots, y_{n-1}\}$ have been obtained, let $y_n \equiv Ay_{n-1}/s_n$ where $s_n$ is the entry of $Ay_{n-1}$ which has largest absolute value. Thus

$$y_n = \frac{AAy_{n-2}}{s_n s_{n-1}} \cdots = \frac{A^n x}{s_n s_{n-1} \cdots s_1} \tag{16.1}$$

$$= \frac{1}{s_n s_{n-1} \cdots s_1} P \begin{pmatrix} J_1^n & \cdots & \vdots \\ & \ddots & \vdots \\ & & J_m^n \end{pmatrix} P^{-1}x$$

$$= \frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} P \begin{pmatrix} \lambda_1^{-n} J_1^n & \cdots & \vdots \\ & \ddots & \vdots \\ & & \lambda_1^{-n} J_m^n \end{pmatrix} P^{-1}x \tag{16.2}$$
Consider one of the blocks in the Jordan form. First consider the \( k^{th} \) of these blocks, \( k > 1 \). It equals
\[
\lambda_1^{-n} J_k^n = \sum_{i=0}^{r_k} \binom{n}{i} \lambda_1^{-n} \lambda_k^{-i} N_k^i
\]
which clearly converges to 0 as \( n \to \infty \) since \( |\lambda_1| > |\lambda_k| \). An application of the ratio test or root test for each term in the sum will show this. When \( k = 1 \), this block is
\[
\lambda_1^{-n} J_1^n = \lambda_1^{-n} J_1^n = \sum_{i=0}^{r_1} \binom{n}{i} \lambda_1^{-n} \lambda_1^{-i} N_1^i = \binom{n}{r_1} [\lambda_1^{-r_1} N_1^r_1 + e_n]
\]
where \( \lim_{n \to \infty} e_n = 0 \) because it is a sum of bounded matrices which are multiplied by \( \binom{n}{r_1} \). This quotient converges to 0 as \( n \to \infty \) because \( i < r_1 \). It follows that \( \binom{r_1}{r_1} \) is of the form
\[
y_n = \frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} \binom{n}{r_1} P \left( \lambda_1^{-r_1} N_1^r_1 + e_n \begin{array}{c} 0 \\ 1 
\end{array} \right) E_n = \frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} \binom{n}{r_1} w_n
\]
where \( E_n \to 0, e_n \to 0 \). Let \( (P^{-1}x)_{m_1} \) denote the first \( m_1 \) entries of the vector \( P^{-1}x \). Unless a very unlucky choice for \( x \) was picked, it will follow that \( (P^{-1}x)_{m_1} \notin \ker (N_1^{r_1}) \). Then for large \( n \), \( y_n \) is close to the vector
\[
\frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} \binom{n}{r_1} P \left( \begin{array}{c} \lambda_1^{-r_1} N_1^r_1 + e_n \\ 0 
\end{array} \right) P^{-1}x = \frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} \binom{n}{r_1} w = z \neq 0
\]
However, this is an eigenvector because
\[
(A - \lambda_1 I) w = P (J - \lambda_1 I) P^{-1} P \left( \begin{array}{c} \lambda_1^{-r_1} N_1^r_1 + e_n \\ 0 
\end{array} \right) P^{-1}x = 0
\]
Recall \( N_1^{r_1+1} = 0 \). Now you could recover an approximation to the eigenvalue as follows.
\[
\frac{(Ay_n, y_n)}{(y_n, y_n)} \approx \frac{(Az, z)}{(z, z)} = \lambda_1
\]
Here \( \approx \) means “approximately equal”. However, there is a more convenient way to identify the eigenvalue in terms of the scaling factors \( s_k \).
\[
\left\| \frac{\lambda_1^n}{s_1 \cdots s_1} \binom{n}{r_1}(w_n - w) \right\|_\infty \approx 0
\]
Pick the largest nonzero entry of \( w, w_i \). Then for large \( n \), it is also likely the case that the largest entry of \( w_n \) will be in the \( l^{th} \) position because \( w_n \) is close to \( w \). From the construction,
In other words, for large $n$

$$\frac{\lambda_1^n}{s_n \cdots s_1} \left(\begin{array}{c} n \\ r_1 \end{array}\right) \approx 1/w_1$$

Therefore, for large $n$,

$$\frac{\lambda_1^n}{s_n \cdots s_1} \left(\begin{array}{c} n \\ r_1 \end{array}\right) \approx \frac{\lambda_1^{n+1}}{s_{n+1} \cdots s_1} \left(\begin{array}{c} n+1 \\ r_1 \end{array}\right)$$

and so

$$\left(\begin{array}{c} n \\ r_1 \end{array}\right)/\left(\begin{array}{c} n+1 \\ r_1 \end{array}\right) \approx \frac{\lambda_1}{s_{n+1}}$$

But $\lim_{n \to \infty} \left(\begin{array}{c} n \\ r_1 \end{array}\right)/\left(\begin{array}{c} n+1 \\ r_1 \end{array}\right) = 1$ and so, for large $n$ it must be the case that $\lambda_1 \approx s_{n+1}$.

This has proved the following theorem which justifies the power method.

**Theorem 16.1.1** Let $A$ be a complex $p \times p$ matrix such that the eigenvalues are $\{\lambda_1, \lambda_2, \cdots, \lambda_r\}$ with $|\lambda_1| > |\lambda_j|$ for all $j \neq 1$. Then for $x$ a given vector, let

$$y_1 = \frac{Ax}{s_1}$$

where $s_1$ is an entry of $Ax$ which has the largest absolute value. If the scalars $\{s_1, \cdots, s_{n-1}\}$ and vectors $\{y_1, \cdots, y_{n-1}\}$ have been obtained, let

$$y_n = \frac{Ay_{n-1}}{s_n}$$

where $s_n$ is the entry of $Ay_{n-1}$ which has largest absolute value. Then it is probably the case that $\{s_n\}$ will converge to $\lambda_1$ and $\{y_n\}$ will converge to an eigenvector associated with $\lambda_1$. If it doesn’t, you picked an incredibly inauspicious initial vector $x$.

In summary, here is the procedure.

**Finding the largest eigenvalue with its eigenvector.**

1. Start with a vector, $u_1$ which you hope is not unlucky.
2. If $u_k$ is known,

$$u_{k+1} = \frac{Au_k}{s_{k+1}}$$

where $s_{k+1}$ is the entry of $Au_k$ which has largest absolute value.
3. When the scaling factors $s_k$ are not changing much, $s_{k+1}$ will be close to the eigenvalue and $u_{k+1}$ will be close to an eigenvector.
4. Check your answer to see if it worked well. If things don’t work well, try another $u_1$. You were miraculously unlucky in your choice.

**Example 16.1.2** Find the largest eigenvalue of $A = \begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix}$. 
You can begin with $\mathbf{u}_1 = (1, \cdots, 1)^T$ and apply the above procedure. However, you can accelerate the process if you begin with $A^n \mathbf{u}_1$ and then divide by the largest entry to get the first approximate eigenvector. Thus

$$
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}^{20} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} =
\begin{pmatrix} 2.555 \times 10^{21} \\ -1.277 \times 10^{21} \\ -3.656 \times 10^{15} \end{pmatrix}
$$

Divide by the largest entry to obtain a good approximation.

$$
\begin{pmatrix}
2.555 \times 10^{21} \\
-1.277 \times 10^{21} \\
-3.656 \times 10^{15}
\end{pmatrix} \frac{1}{2.555 \times 10^{21}} =
\begin{pmatrix}
1.0 \\
-0.5 \\
-1.430 \times 10^{-6}
\end{pmatrix}
$$

Now begin with this one.

$$
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix} \begin{pmatrix} 1.0 \\ -0.5 \\ -1.430 \times 10^{-6} \end{pmatrix} =
\begin{pmatrix} 12.000 \\ -6.000 \times 10^{-6} \\ 4.291 \times 10^{-6} \end{pmatrix}
$$

Divide by 12 to get the next iterate.

$$
\begin{pmatrix}
12.000 \\
-6.000 \times 10^{-6} \\
4.291 \times 10^{-6}
\end{pmatrix} \frac{1}{12} =
\begin{pmatrix}
1.0 \\
-0.5 \\
3.576 \times 10^{-7}
\end{pmatrix}
$$

Another iteration will reveal that the scaling factor is still 12. Thus this is an approximate eigenvalue. In fact, it is the largest eigenvalue and the corresponding eigenvector is $\begin{pmatrix} 1.0 & -0.5 & 0 \end{pmatrix}$. The process has worked very well.

### 16.1.1 The Shifted Inverse Power Method

This method can find various eigenvalues and eigenvectors. It is a significant generalization of the above simple procedure and yields very good results. One can find complex eigenvalues using this method. The situation is this: You have a number $\alpha$ which is close to $\lambda$, some eigenvalue of an $n \times n$ matrix $A$. You don’t know $\lambda$ but you know that $\alpha$ is closer to $\lambda$ than to any other eigenvalue. Your problem is to find both $\lambda$ and an eigenvector which goes with $\alpha$. Another way to look at this is to start with $\alpha$ and seek the eigenvalue $\lambda$, which is closest to $\alpha$ along with an eigenvector associated with $\lambda$. If $\alpha$ is an eigenvalue of $A$, then you have what you want. Therefore, I will always assume $\alpha$ is not an eigenvalue of $A$ and so $(A - \alpha I)^{-1}$ exists. The method is based on the following lemma.

**Lemma 16.1.3** Let $\{\lambda_k\}_{k=1}^{n}$ be the eigenvalues of $A$. If $x_k$ is an eigenvector of $A$ for the eigenvalue $\lambda_k$, then $x_k$ is an eigenvector for $(A - \alpha I)^{-1}$ corresponding to the eigenvalue $\frac{1}{\lambda_k - \alpha}$. Conversely, if

$$
(A - \alpha I)^{-1} y = \frac{1}{\lambda - \alpha} y
$$

and $y \neq 0$, then $Ay = \lambda y$.

**Proof:** Let $\lambda_k$ and $x_k$ be as described in the statement of the lemma. Then

$$(A - \alpha I) x_k = (\lambda_k - \alpha) x_k$$

and so

$$
\frac{1}{\lambda_k - \alpha} x_k = (A - \alpha I)^{-1} x_k.
$$
Suppose $y = \frac{1}{\lambda - \alpha} [Ay - \alpha y]$. Solving for $Ay$ leads to $Ay = \lambda y$. ■

Now assume $\alpha$ is closer to $\lambda$ than to any other eigenvalue. Then the magnitude of $\frac{1}{\lambda - \alpha}$ is greater than the magnitude of all the other eigenvalues of $(A - \alpha I)^{-1}$. Therefore, the power method applied to $(A - \alpha I)^{-1}$ will yield $\frac{1}{\lambda - \alpha}$. You end up with $s_{n+1} \approx \frac{1}{\lambda - \alpha}$ and solve for $\lambda$.

### 16.1.2 The Explicit Description Of The Method

Here is how you use this method to find the eigenvalue closest to $\alpha$ and the corresponding eigenvector.

1. Find $(A - \alpha I)^{-1}$.

2. Pick $u_1$. If you are not phenomenally unlucky, the iterations will converge.

3. If $u_k$ has been obtained,

$$ u_{k+1} = \frac{(A - \alpha I)^{-1} u_k}{s_{k+1}} $$

where $s_{k+1}$ is the entry of $(A - \alpha I)^{-1} u_k$ which has largest absolute value.

4. When the scaling factors, $s_k$ are not changing much and the $u_k$ are not changing much, find the approximation to the eigenvalue by solving

$$ s_{k+1} = \frac{1}{\lambda - \alpha} $$

for $\lambda$. The eigenvector is approximated by $u_{k+1}$.

5. Check your work by multiplying by the original matrix to see how well what you have found works.

Thus this amounts to the power method for the matrix $(A - \alpha I)^{-1}$ but you are free to pick $\alpha$.

### 16.2 Automation With Matlab

You can do the above example and other examples using Matlab. Here are some commands which will do this. It is done here for a $3 \times 3$ matrix but you adapt for any size.

```matlab
a=[5 -8 6;1 0 0;0 1 0]; b=i; F=inv(a-b*eye(3));
S=1; u=[1;1;1]; d=1; k=1;
while d >.00001 & k<1000
w=F*u; [M,I]=max(abs(w)); T=w(I); u=w/T;
d=abs(T-S); S=T; k=k+1;
end
u
b+1/T
k
a*u-(b+1/T)*u
```

`eye(3)` signifies the $3 \times 3$ identity. It is less trouble to write this.

Note how the “while loop” is limited to 1000 iterations. That way it won’t go on forever if there is something wrong. This asks for the eigenvalue closest to $\alpha = i$. When Matlab stalls, to get it to quit, you type control c. The last line checks the answer and the line with $k$ tells the number of iterations used. Also, the funny notation $[M,I]=\max(abs(w))$: $T=w(I)$; gets it to pick out the entry which has largest absolute value $w(I)$ and keep that entry unchanged. The above iteration finds
the eigenvalue closest to \( i \) along with the corresponding eigenvector. When the procedure does not work well for \( b \) real, you might imagine that there are complex eigenvalues and so, since the above procedure is going to give you real approximations, it can’t find the complex eigenvalues. Thus you should take \( b \) to be complex as done above.

If you have Matlab work the above iteration, you get the following for the eigenvector eigenvalue and number of iterations, and error .

\[
\begin{pmatrix}
1 \\
0.5 - 0.5i \\
-0.5i
\end{pmatrix}
, \quad 1 + i, \quad k = 18, \quad 10^{-5}

\begin{pmatrix}
0 \\
-0.1321 + 0.1862i \\
-0.1325 + 0.1863i
\end{pmatrix}
\]

In fact, this eigenvector is exactly right as is the eigenvalue \( 1 + i \).

Thus this method will find eigenvalues real or complex along with an eigenvector associated with the eigenvalue. Note that the characteristic polynomial of the above matrix is \( \lambda^3 - 5\lambda^2 + 8\lambda - 6 \) and the above finds a complex root to this polynomial. More generally, if you have a polynomial \( \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0 \), a matrix which has this as its characteristic polynomial is called a companion matrix and you can show a matrix which works for this polynomial is of the form

\[
\begin{pmatrix}
-a_{n-1} & -a_{n-2} & \cdots & a_0 \\
1 & 0 & & \\
& \ddots & \ddots & \\
& & 1 & 0
\end{pmatrix}
\]

Thus this method is capable of finding roots to a polynomial equation which are close to a given complex number. Of course there is a problem with determining which number you should pick. A way to determine this will be discussed later. It involves something called the QR algorithm.

**Example 16.2.1** Find the eigenvalue of

\[
A = \begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\]

which is closest to \(-7\). Also find an eigenvector which goes with this eigenvalue.

We use the algorithm described above.

\[
a = [5 \ -14 \ 11; \ 4 \ -4; \ 3 \ 6 \ -3]; \ b = -7; \ F = \text{inv}(a - b*\text{eye}(3)); \\
S = 1; \ u = [1; 1; 1]; \ d = 1; \ k = 1; \\
\text{while} \ d > 0.001 \ & \ k < 1000 \\
w = F*u; \ [M, I] = \text{max}(\text{abs}(w)); \ T = w(I); \ u = w/T; \\
d = \text{abs}(T-S); \ S = T; \ k = k+1; \\
\text{end} \\
u \ \\
k \ \\
b+1/T \\
a*u-(b+1/T)*u
\]

This yields the following after 8 iterations.

\[
\begin{pmatrix}
1 \\
0 \\
-1
\end{pmatrix}
, \quad -6
\]

for the eigenvector and eigenvalue. In fact, this is exactly correct.
Example 16.2.2 Consider the symmetric matrix \( A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix} \). Find the middle eigenvalue and an eigenvector which goes with it.

Since \( A \) is symmetric, it follows it has three real eigenvalues which are solutions to

\[
p(\lambda) = \det \left( \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix} \right) = \lambda^3 - 4\lambda^2 - 24\lambda - 17 = 0
\]

If you use your graphing calculator to graph this polynomial, you find there is an eigenvalue somewhere between \(-.9\) and \(-.8\) and that this is the middle eigenvalue. Using \(-.8\) as the number close to the eigenvalue desired, after 7 iterations, you get

\[
u = \begin{pmatrix} 1 \\ -0.5878 \\ -0.2271 \end{pmatrix}, \quad \lambda = -0.8569
\]

Note that

\[
\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix} - (-0.8569) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 2.5244 \times 10^{-29} \\ 1.1418 \times 10^{-4} \\ -1.99 \times 10^{-6} \end{pmatrix}
\]

There is an easy to use trick which will eliminate some of the fuss and bother in using the shifted inverse power method. If you have \((A - \alpha I)^{-1} x = \mu x\) then multiplying through by \((A - \alpha I)\), one finds that \( x \) will be an eigenvector for \( A \) with eigenvalue \( \alpha + \mu^{-1} \). Hence you could simply take \((A - \alpha I)^{-1}\) to a high power and multiply by a vector to get a vector which points in the direction of an eigenvalue of \( A \). Then divide by the largest entry and identify the eigenvalue directly by multiplying the eigenvector by \( A \). This is illustrated in the next example.

Example 16.2.3 Find the eigenvalue near \(-1.2\) along with an eigenvector.

\[
A = \begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix}.
\]

This is only a 3\(\times\)3 matrix and so it is not hard to estimate the eigenvalues. Just get the characteristic equation, graph it using a calculator and zoom in to find the eigenvalues. If you do this, you find there is an eigenvalue near \(-1.2\), one near \(-.4\), and one near 5.5. (The characteristic equation is \(2 + 8\lambda + 4\lambda^2 - \lambda^3 = 0\).) Of course we have no idea what the eigenvectors are.

Lets first try to find the eigenvector and an approximation for the eigenvalue near \(-1.2\). In this case, let \( \alpha = -1.2 \). Then

\[
(A - \alpha I)^{-1} = \begin{pmatrix} -25.357143 & -33.928571 & 50.0 \\ 12.5 & 17.5 & -25.0 \\ 23.214286 & 30.357143 & -45.0 \end{pmatrix}.
\]
Then

\[
\begin{pmatrix}
-25.357143 & -33.928571 & 50.0 \\
12.5 & 17.5 & -25.0 \\
23.214286 & 30.357143 & -45.0
\end{pmatrix}^{17}
\begin{pmatrix}
1 \\
1 \\
1
\end{pmatrix}
\]

\[
= \begin{pmatrix}
-4.9432 \times 10^{28} \\
2.4312 \times 10^{28} \\
4.4928 \times 10^{28}
\end{pmatrix}
\]

The initial approximation for an eigenvector will then be the above divided by its largest entry.

\[
\begin{pmatrix}
-4.9432 \times 10^{28} \\
2.4312 \times 10^{28} \\
4.4928 \times 10^{28}
\end{pmatrix}
\frac{1}{-4.9432 \times 10^{28}}
\begin{pmatrix}
1 \\
0 \\
0
\end{pmatrix}
= \begin{pmatrix}
1.0 \\
-0.49183 \\
-0.90888
\end{pmatrix}
\]

How close is this to being an eigenvector?

\[
\begin{pmatrix}
2 & 1 & 3 \\
2 & 1 & 1 \\
3 & 2 & 1
\end{pmatrix}
\begin{pmatrix}
1.0 \\
-0.49183 \\
-0.90888
\end{pmatrix}
= \begin{pmatrix}
-1.2185 \\
0.59929 \\
1.1075
\end{pmatrix}
\]

\[
\begin{pmatrix}
2 & 1 & 3 \\
2 & 1 & 1 \\
3 & 2 & 1
\end{pmatrix}
\begin{pmatrix}
1.0 \\
-0.49183 \\
-0.90888
\end{pmatrix}
= \begin{pmatrix}
-1.2185 \\
0.59929 \\
1.1075
\end{pmatrix}
\]

For all practical purposes, this has found the eigenvector and eigenvalue of \(-1.2185\).

### 16.2.1 Complex Eigenvalues

What about complex eigenvalues? If your matrix is real, you won’t see these by graphing the characteristic equation on your calculator. Will the shifted inverse power method find these eigenvalues and their associated eigenvectors? The answer is yes. However, for a real matrix, you must pick \(\alpha\) to be complex. This is because the eigenvalues occur in conjugate pairs so if you don’t pick it complex, it will be the same distance between any conjugate pair of complex numbers and so nothing in the above argument for convergence implies you will get convergence to a complex number. Also, the process of iteration will yield only real vectors and scalars.

**Example 16.2.4** Find the complex eigenvalues and corresponding eigenvectors for the matrix

\[
\begin{pmatrix}
5 & -8 & 6 \\
1 & 0 & 0 \\
0 & 1 & 0
\end{pmatrix}
\]

Here the characteristic equation is \(\lambda^3 - 5\lambda^2 + 8\lambda - 6 = 0\). One solution is \(\lambda = 3\). The other two are \(1 + i\) and \(1 - i\). I will apply the process to \(\alpha = i\) to find the eigenvalue closest to \(i\). The above algorithm yields the following after 15 iterations.

\[
u = \begin{pmatrix}
1 \\
.5 - .5i \\
-.5i
\end{pmatrix}, \quad \lambda = 1 + i
\]
This illustrates an interesting topic which leads to many related topics. If you have a polynomial, $x^4 + ax^3 + bx^2 + cx + d$, you can consider it as the characteristic polynomial of a certain matrix, called a companion matrix. In this case,

$$
\begin{pmatrix}
-a & -b & -c & -d \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0
\end{pmatrix}.
$$

The above example was just a companion matrix for $\lambda^3 - 5\lambda^2 + 8\lambda - 6$. You can see the pattern which will enable you to obtain a companion matrix for any polynomial of the form $\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n$. This illustrates that one way to find the complex zeros of a polynomial is to use the shifted inverse power method on a companion matrix for the polynomial. Doubtless there are better ways but this does illustrate how impressive this procedure is. Do you have a better way?

Note that the shifted inverse power method is a way you can begin with something close but not equal to an eigenvalue and end up with something close to an eigenvector.

### 16.2.2 Rayleigh Quotients And Estimates for Eigenvalues

There are many specialized results concerning the eigenvalues and eigenvectors for Hermitian matrices. Recall a matrix $A$ is Hermitian if $A = A^*$ where $A^*$ means to take the transpose of the conjugate of $A$. In the case of a real matrix, Hermitian reduces to symmetric. Recall also that for $x \in \mathbb{F}^n$,

$$|x|^2 = x^*x = \sum_{j=1}^{n} |x_j|^2.$$

Recall the following corollary found on Page 311 which is stated here for convenience.

**Corollary 16.2.5** If $A$ is Hermitian, then all the eigenvalues of $A$ are real and there exists an orthonormal basis of eigenvectors.

Thus for $\{\mathbf{x}_k\}_{k=1}^n$ this orthonormal basis,

$$x_i^*x_j = \delta_{ij} \equiv \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

For $x \in \mathbb{F}^n$, $x \neq 0$, the Rayleigh quotient is defined by

$$\frac{x^*Ax}{|x|^2}.$$

Now let the eigenvalues of $A$ be $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and $Ax_k = \lambda_kx_k$ where $\{x_k\}_{k=1}^n$ is the above orthonormal basis of eigenvectors mentioned in the corollary. Then if $x$ is an arbitrary vector, there exist constants, $a_i$ such that

$$x = \sum_{i=1}^{n} a_i x_i.$$

Also,

$$|x|^2 = \sum_{i=1}^{n} |x_i|^2 \sum_{j=1}^{n} a_jx_j = \sum_{ij} \pi_i a_j x_i^*x_j = \sum_{ij} \pi_i a_j \delta_{ij} = \sum_{i=1}^{n} |a_i|^2.$$

Therefore,

$$\frac{x^*Ax}{|x|^2} = \frac{\left(\sum_{i=1}^{n} |x_i|^2\right) \left(\sum_{j=1}^{n} a_j \lambda_j x_j\right)}{\sum_{i=1}^{n} |a_i|^2} = \frac{\sum_{ij} \pi_i a_j \lambda_j x_i^*x_j}{\sum_{i=1}^{n} |a_i|^2}$$
In other words, the Rayleigh quotient is always between the largest and the smallest eigenvalues of \( A \). When \( x = x_n \), the Rayleigh quotient equals the largest eigenvalue and when \( x = x_1 \) the Rayleigh quotient equals the smallest eigenvalue. Suppose you calculate a Rayleigh quotient. How close is it to some eigenvalue?

**Theorem 16.2.6** Let \( x \neq 0 \) and form the Rayleigh quotient,

\[
\frac{x^*Ax}{|x|^2} = q.
\]

Then there exists an eigenvalue of \( A \), denoted here by \( \lambda_q \) such that

\[
|\lambda_q - q| \leq \frac{|Ax - qx|}{|x|}.
\]

**Proof:** Let \( x = \sum_{k=1}^n a_kx_k \) where \( \{x_k\}_{k=1}^n \) is the orthonormal basis of eigenvectors.

\[
|Ax - qx|^2 = (Ax - qx)^* (Ax - qx)
\]

\[
= \left( \sum_{k=1}^n a_k\lambda_k x_k - qa_k x_k \right)^* \left( \sum_{k=1}^n a_k\lambda_k x_k - qa_k x_k \right)
\]

\[
= \left( \sum_{j=1}^n (\lambda_j - q)\bar{a}_j x_j^* \right) \left( \sum_{k=1}^n (\lambda_k - q) a_k x_k \right)
\]

\[
= \sum_{j,k} (\lambda_j - q)\bar{a}_j (\lambda_k - q) a_k x_j^* x_k
\]

Now pick the eigenvalue \( \lambda_q \) which is closest to \( q \). Then

\[
|Ax - qx|^2 = \sum_{k=1}^n |a_k|^2 (\lambda_k - q)^2 \geq (\lambda_q - q)^2 \sum_{k=1}^n |a_k|^2 = (\lambda_q - q)^2 |x|^2
\]

which implies \( 16.4 \).

**Example 16.2.7** Consider the symmetric matrix \( A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix} \). Let \( x = (1, 1, 1)^T \). How close is the Rayleigh quotient to some eigenvalue of \( A \)? Find the eigenvector and eigenvalue to several decimal places.

Everything is real and so there is no need to worry about taking conjugates. Therefore, the Rayleigh quotient is

\[
\begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \frac{19}{3}
\]
According to the above theorem, there is some eigenvalue of this matrix $\lambda_q$ such that

$$
\left| \lambda_q - \frac{19}{3} \right| \leq \left| \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix} - \frac{19}{3} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \right| = \frac{1}{\sqrt{3}} \begin{pmatrix} -\frac{1}{3} \\ \frac{4}{3} \\ \frac{4}{9} \end{pmatrix} = \sqrt{\frac{1}{9} + \left(\frac{4}{3}\right)^2 + \left(\frac{4}{9}\right)^2} = 1.2472
$$

Could you find this eigenvalue and associated eigenvector? Of course you could. This is what the shifted inverse power method is all about.

Solve

$$
\begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix} - \frac{19}{3} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}
$$

In other words solve

$$
\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}
$$

and divide by the entry which is largest, 3.8707, to get

$$
u_2 = \begin{pmatrix} .69925 \\ .49389 \\ 1.0 \end{pmatrix}
$$

Now solve

$$
\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} .69925 \\ .49389 \\ 1.0 \end{pmatrix}
$$

and divide by the largest entry, 2.9979 to get

$$
u_3 = \begin{pmatrix} .71473 \\ .52263 \\ 1.0 \end{pmatrix}
$$

Now solve

$$
\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} .71473 \\ .52263 \\ 1.0 \end{pmatrix}
$$

and divide by the largest entry, 3.0454, to get

$$
u_4 = \begin{pmatrix} .7137 \\ .52056 \\ 1.0 \end{pmatrix}
$$

Solve

$$
\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} .7137 \\ .52056 \\ 1.0 \end{pmatrix}
$$
and divide by the largest entry, 3.0421 to get

\[ u_5 = \begin{pmatrix} \frac{1.71378}{7.1378} \\ \frac{0.52073}{1.0} \end{pmatrix} \]

You can see these scaling factors are not changing much. The predicted eigenvalue is then about

\[ \frac{1}{3.0421} + \frac{19}{3} \approx 6.6621. \]

How close is this?

\[
\begin{pmatrix}
1 & 2 & 3 \\
2 & 2 & 1 \\
3 & 1 & 4
\end{pmatrix}
\begin{pmatrix}
\frac{1.71378}{7.1378} \\
\frac{0.52073}{1.0}
\end{pmatrix}
= \begin{pmatrix} 4.7552 \\ 3.469 \end{pmatrix}
\]

while

\[ 6.6621 \begin{pmatrix}
\frac{1.71378}{7.1378} \\
\frac{0.52073}{1.0}
\end{pmatrix}
= \begin{pmatrix} 4.7553 \\ 3.4692 \end{pmatrix}. \]

You see that for practical purposes, this has found the eigenvalue and an eigenvector.

### 16.3 The QR Algorithm

#### 16.3.1 Basic Properties And Definition

Recall the theorem about the QR factorization in Theorem 13.2.9. It says that given an \( n \times n \) real matrix \( A \), there exists a real orthogonal matrix \( Q \) and an upper triangular matrix \( R \) such that \( A = QR \) and that this factorization can be accomplished by a systematic procedure. One such procedure was given in proving this theorem.

**Theorem 16.3.1** Let \( A \) be an \( n \times n \) complex matrix. Then there exists a unitary \( Q \) and upper triangular \( R \) such that \( A = QR \).

**Proof:** This is obvious if \( n = 1 \). Suppose true for \( n \) and let

\[ A = \begin{pmatrix}
a_1 & \cdots & a_n & a_{n+1}
\end{pmatrix} \]

Let \( Q_1 \) be a unitary matrix such that \( Q_1 a_1 = |a_1| e_1 \) in case \( a_1 \neq 0 \). If \( a_1 = 0 \), let \( Q_1 = I \). Thus

\[ Q_1 A = \begin{pmatrix} a & b \\ 0 & A_1 \end{pmatrix} \]

where \( A_1 \) is \((n - 1) \times (n - 1)\). By induction, there exists \( Q_2 \) an \((n - 1) \times (n - 1)\) unitary matrix such that \( Q_2 A_1 = R' \), an upper triangular matrix. Then

\[ \begin{pmatrix} 1 & 0 \\ 0 & Q_2 \end{pmatrix} Q_1 A = \begin{pmatrix} a & b \\ 0 & R' \end{pmatrix} = R \]

Since the product of unitary matrices is unitary, there exists \( Q \) unitary such that \( Q^* A = R \) and so \( A = QR \).

The QR algorithm is described in the following definition.
**Definition 16.3.2** The QR algorithm is the following. In the description of this algorithm, Q is unitary and R is upper triangular having nonnegative entries on the main diagonal. Starting with $A$ an $n \times n$ matrix, form

$$A_0 = A = Q_1 R_1$$

Then

$$A_1 = R_1 Q_1.$$  \hfill (16.5)

In general given

$$A_k = R_k Q_k, \quad A_{k+1} = R_{k+1} Q_{k+1}$$

obtain $A_{k+1}$ by

$$A_k = Q_{k+1} R_{k+1}, \quad A_{k+1} = R_{k+1} Q_{k+1}$$ \hfill (16.6)

This algorithm was proposed by Francis in 1961. The sequence $\{A_k\}$ is the desired sequence of iterates. Now with the above definition of the algorithm, here are its properties. The next lemma shows each of the $A_k$ is unitarily similar to $A$ and the amazing thing about this algorithm is that often it becomes increasingly easy to find the eigenvalues of the $A_k$.

**Lemma 16.3.3** Let $A$ be an $n \times n$ matrix and let the $Q_k$ and $R_k$ be as described in the algorithm. Then each $A_k$ is unitarily similar to $A$ and denoting by $Q^{(k)}$ the product $Q_1 Q_2 \cdots Q_k$ and $R^{(k)}$ the product $R_k R_{k-1} \cdots R_1$, it follows that

$$A^k = Q^{(k)} R^{(k)}$$

(The matrix on the left is $A$ raised to the $k$th power.)

$$A = Q^{(k)} A_k Q^{(k)*}, \quad A_k = Q^{(k)*} A Q^{(k)}.$$  \hfill (16.7)

**Proof:** From the algorithm, $R_{k+1} = A_{k+1} Q_{k+1}^*$ and so

$$A_k = Q_{k+1} R_{k+1} = Q_{k+1} A_{k+1} Q_{k+1}^*.$$  \hfill (16.8)

Now iterating this, it follows

$$A_{k-1} = Q_k A_k Q_k^* = Q_k Q_{k+1} A_{k+1} Q_{k+1}^* Q_k^*,$$

$$A_{k-2} = Q_{k-1} A_{k-1} Q_{k-1}^* = Q_{k-1} Q_k Q_{k+1} A_{k+1} Q_{k+1}^* Q_k^* Q_{k-1}^*.$$  

etc. Thus, after $k - 2$ more iterations,

$$A = Q^{(k+1)} A_{k+1} Q^{(k+1)*}.$$  

The product of unitary matrices is unitary and so this proves the first claim of the lemma.

Now consider the part about $A^k$. From the algorithm, this is clearly true for $k = 1$. ($A^1 = QR$) Suppose then that

$$A^k = Q_1 Q_2 \cdots Q_k R_k R_{k-1} \cdots R_1$$

What was just shown indicated

$$A = Q_1 Q_2 \cdots Q_{k+1} A_{k+1} Q_{k+1}^* Q_k^* \cdots Q_1^*$$

and now from the algorithm, $A_{k+1} = R_{k+1} Q_{k+1}$ and so

$$A = Q_1 Q_2 \cdots Q_{k+1} R_{k+1} Q_{k+1}^* Q_k^* \cdots Q_1^*.$$  

Then

$$A^{k+1} = AA^k =$$

$$Q_1 Q_2 \cdots Q_{k+1} R_{k+1} Q_{k+1}^* Q_k^* \cdots Q_1^* Q_1 R_k R_{k-1} \cdots R_1$$

$$= Q_1 Q_2 \cdots Q_{k+1} R_{k+1} R_k R_{k-1} \cdots R_1 = Q^{(k+1)} R^{(k+1)} \quad \blacksquare$$

Here is another very interesting lemma.
Lemma 16.3.4 Suppose $Q^{(k)}, Q$ are unitary and $R_k$ is upper triangular such that the diagonal entries on $R_k$ are all positive and

$$Q = \lim_{k \to \infty} Q^{(k)} R_k$$

Then

$$\lim_{k \to \infty} Q^{(k)} = Q, \quad \lim_{k \to \infty} R_k = I.$$ 

Also the $QR$ factorization of $A$ is unique whenever $A^{-1}$ exists.

**Proof:** Let

$$Q = (q_1, \ldots, q_n), \quad Q^{(k)} = (q_1^k, \ldots, q_n^k)$$

where the $q$ are the columns. Also denote by $r_{ij}^k$ the $ij^{th}$ entry of $R_k$. Thus

$$Q^{(k)} R_k = (q_1^k, \ldots, q_n^k) \begin{pmatrix} r_{11}^k & * \\ \vdots \\ 0 & r_{nn}^k \end{pmatrix}$$

It follows

$$r_{11}^k q_1^k \to q_1,$n$$

and so

$$r_{11}^k q_1^k \to q_1.$$ 

Therefore,

$$q_1^k \to q_1.$$ 

Next consider the second column.

$$r_{12}^k q_1^k + r_{22}^k q_2^k \to q_2$$

Taking the inner product of both sides with $q_1^k$ it follows

$$\lim_{k \to \infty} r_{12}^k = \lim_{k \to \infty} (q_2 \cdot q_1^k) = (q_2 \cdot q_1) = 0.$$ 

Therefore,

$$\lim_{k \to \infty} r_{22}^k q_2^k = q_2$$

and since $r_{22}^k > 0$, it follows as in the first part that $r_{22}^k \to 1$. Hence

$$\lim_{k \to \infty} q_2^k = q_2.$$ 

Continuing this way, it follows

$$\lim_{k \to \infty} r_{ij}^k = 0$$

for all $i \neq j$ and

$$\lim_{k \to \infty} r_{ij}^k = 1, \quad \lim_{k \to \infty} q_j^k = q_j.$$ 

Thus $R_k \to I$ and $Q^{(k)} \to Q$. This proves the first part of the lemma.

The second part follows immediately. If $QR = Q'R' = A$ where $A^{-1}$ exists, then

$$Q^* Q' = R (R')^{-1}$$

and I need to show both sides of the above are equal to $I$. The left side of the above is unitary and the right side is upper triangular having positive entries on the diagonal. This is because the inverse of such an upper triangular matrix having positive entries on the main diagonal is still upper triangular having positive entries on the main diagonal and the product of two such upper triangular matrices gives another of the same form having positive entries on the main diagonal. Suppose then
that \( Q = R \) where \( Q \) is unitary and \( R \) is upper triangular having positive entries on the main diagonal. Let \( Q_k = Q \) and \( R_k = R \). It follows

\[
IR_k \rightarrow R = Q
\]

and so from the first part, \( R_k \rightarrow I \) but \( R_k = R \) and so \( R = I \). Thus applying this to \( Q^*Q' = R(R')^{-1} \) yields both sides equal \( I \).

A case of all this is of great interest. Suppose \( A \) has a largest eigenvalue \( \lambda \) which is real. Then \( A^n \) is of the form \( (A^{n-1}a_1, \ldots, A^{n-1}a_n) \) and so likely each of these columns will be pointing roughly in the direction of an eigenvector of \( A \) which corresponds to this eigenvalue. Then when you do the QR factorization of this, it follows from the fact that \( R \) is upper triangular, that the first column of \( Q \) will be a multiple of \( A^{n-1}a_1 \) and so will end up being roughly parallel to the eigenvector desired. Also this will require the entries below the top in the first column of \( A_n = Q^T A Q \) will all be small because they will be of the form \( q_i^T A q_1 \approx \lambda q_i^T q_1 = 0 \). Therefore, \( A_n \) will be of the form

\[
\begin{pmatrix}
\lambda' & a \\
e & B
\end{pmatrix}
\]

where \( e \) is small. It follows that \( \lambda' \) will be close to \( \lambda \) and \( q_1 \) will be close to an eigenvector for \( \lambda \). Then if you like, you could do the same thing with the matrix \( B \) to obtain approximations for the other eigenvalues. Finally, you could use the shifted inverse power method to get more exact solutions.

### 16.3.2 The Case Of Real Eigenvalues

With these lemmas, it is possible to prove that for the QR algorithm and certain conditions, the sequence \( A_k \) converges pointwise to an upper triangular matrix having the eigenvalues of \( A \) down the diagonal. I will assume all the matrices are real here.

This convergence won’t always happen. Consider for example the matrix

\[
\begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}
\]

You can verify quickly that the algorithm will return this matrix for each \( k \). The problem here is that, although the matrix has the two eigenvalues \(-1, 1\), they have the same absolute value. The QR algorithm works in somewhat the same way as the power method, exploiting differences in the size of the eigenvalues.

If \( A \) has all real eigenvalues and you are interested in finding these eigenvalues along with the corresponding eigenvectors, you could always consider \( A + \lambda I \) instead where \( \lambda \) is sufficiently large and positive that \( A + \lambda I \) has all positive eigenvalues. (Recall Gerschgorin’s theorem.) Then if \( \mu \) is an eigenvalue of \( A + \lambda I \) with

\[
(A + \lambda I)x = \mu x
\]

then

\[
Ax = (\mu - \lambda)x
\]

so to find the eigenvalues of \( A \) you just subtract \( \lambda \) from the eigenvalues of \( A + \lambda I \). Thus there is no loss of generality in assuming at the outset that the eigenvalues of \( A \) are all positive. Here is the theorem. It involves a technical condition which will often hold. The proof presented here follows [52] and is a special case of that presented in this reference.

Before giving the proof, note that the product of upper triangular matrices is upper triangular. If they both have positive entries on the main diagonal so will the product. Furthermore, the inverse of an upper triangular matrix is upper triangular. I will use these simple facts without much comment whenever convenient.

**Theorem 16.3.5** Let \( A \) be a real matrix having eigenvalues

\[
\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0
\]
and let
\[ A = SDS^{-1} \]  \hspace{1cm} (16.9)
where
\[ D = \begin{pmatrix} 
\lambda_1 & 0 \\
& \ddots \\
n & \lambda_n 
\end{pmatrix} \]
and suppose \( S^{-1} \) has an \( LU \) factorization. Then the matrices \( A_k \) in the \( QR \) algorithm described above converge to an upper triangular matrix \( T' \) having the eigenvalues of \( A; \lambda_1, \ldots, \lambda_n \) descending on the main diagonal. The matrices \( Q^{(k)} \) converge to \( Q' \), an orthogonal matrix which equals \( Q \) except for possibly having some columns multiplied by \(-1\) for \( Q \) the unitary part of the \( QR \) factorization of \( S \),
\[ S = QR, \]
and
\[ \lim_{k \to \infty} A_k = T' = Q'^T AQ' \]

**Proof:** From Lemma \[16.3.3\]
\[ A^k = Q^{(k)}R^{(k)} = SD^kS^{-1} \]  \hspace{1cm} (16.10)
Let \( S = QR \) where this is just a \( QR \) factorization which is known to exist and let \( S^{-1} = LU \) which is assumed to exist. Thus
\[ Q^{(k)}R^{(k)} = QRD^kLU \]  \hspace{1cm} (16.11)
and so
\[ Q^{(k)}R^{(k)} = QRD^kLU = QRD^kLD^{-k}D^kU \]
That matrix in the middle, \( D^kLD^{-k} \) satisfies
\[ (D^kLD^{-k})_{ij} = \lambda_i^kL_{ij}\lambda_j^{-k} \text{ for } j \leq i, \ 0 \text{ if } j > i. \]
Thus for \( j < i \) the expression converges to 0 because \( \lambda_j > \lambda_i \) when this happens. When \( i = j \) it reduces to 1. Thus the matrix in the middle is of the form \( I + E_k \) where \( E_k \to 0 \). Then it follows
\[ A^k = Q^{(k)}R^{(k)} = QR(I + E_k)D^kU \]
\[ = Q(I + RE_kR^{-1})RD^kU = Q(I + F_k)RD^kU \]
where \( F_k \to 0 \). Then let \( I + F_k = Q_kR_k \) where this is another \( QR \) factorization. Then it reduces to
\[ Q^{(k)}R^{(k)} = QQ_kR_kRD^kU \]
This looks really interesting because by Lemma \[16.3.4\] \( Q_k \to I \) and \( R_k \to I \) because \( Q_kR_k = (I + F_k) \to I \). So it follows \( QQ_k \) is an orthogonal matrix converging to \( Q \) while
\[ R_kRD^kU \left(R^{(k)}\right)^{-1} \]
is upper triangular, being the product of upper triangular matrices. Unfortunately, it is not known that the diagonal entries of this matrix are nonnegative because of the \( U \). Let \( \Lambda \) be just like the identity matrix but having some of the ones replaced with \(-1\) in such a way that \( \Lambda U \) is an upper triangular matrix having positive diagonal entries. Note \( \Lambda^2 = I \) and also \( \Lambda \) commutes with a diagonal matrix. Thus
\[ Q^{(k)}R^{(k)} = QQ_kR_kRD^k\Lambda^2U = QQ_kR_kR\Lambda D^k(\Lambda U) \]
At this point, one does some inspired massaging to write the above in the form
\[ QQ_k \left( AD^k \right) \left[ (AD^k)^{-1} R_k R A D^k \right] (AU) \]
\[ = Q (Q_k A) D^k \left[ (AD^k)^{-1} R_k R A D^k \right] (AU) \]
\[ = Q (Q_k A) D^k \left[ (AD^k)^{-1} R_k R A D^k \right] (AU) \]
\[ \equiv G_k \]
Now I claim the middle matrix in \([\cdot]\) is upper triangular and has all positive entries on the diagonal. This is because it is an upper triangular matrix which is similar to the upper triangular matrix \( R_k R \) and so it has the same eigenvalues (diagonal entries) as \( R_k R \). Thus the matrix \( G_k \equiv D^k \left[ (AD^k)^{-1} R_k R A D^k \right] (AU) \) is upper triangular and has all positive entries on the diagonal. Multiply on the right by \( G_k^{-1} \) to get
\[ Q^{(k)} R^{(k)} G_k^{-1} = Q Q_k A \rightarrow Q' \]
where \( Q' \) is essentially equal to \( Q \) but might have some of the columns multiplied by \(-1\). This is because \( Q_k \rightarrow I \) and so \( Q_k A \rightarrow A \). Now by Lemma 16.3.3 it follows
\[ Q^{(k)} \rightarrow Q', \ R^{(k)} G_k^{-1} \rightarrow I. \]
It remains to verify \( A_k \) converges to an upper triangular matrix. Recall that from [QR] and the definition below this \( (S = QR) \)
\[ A = SDS^{-1} = (QR) D (QR)^{-1} = Q R D R^{-1} Q^T = QTQ^T \]
Where \( T \) is an upper triangular matrix. This is because it is the product of upper triangular matrices \( R, D, R^{-1} \). Thus \( Q^T AQ = T \). If you replace \( Q \) with \( Q' \) in the above, it still results in an upper triangular matrix \( T' \) having the same diagonal entries as \( T \). This is because
\[ T = Q^T AQ = (Q'A)^T A (Q'A) = AQ'^T AQ'A \]
and considering the \( ij \) entry yields
\[ (Q^T AQ)_{ij} = \sum_{j,k} A_{ij} (Q^T AQ')_{jk} A_{ki} = A_{ii} A_{ii} (Q^T AQ')_{ii} = (Q^T AQ')_{ii} \]
Recall from Lemma 16.3.3, \( A_k = Q^{(k)} T AQ^{(k)} \). Thus taking a limit and using the first part,
\[ A_k = Q^{(k)} T AQ^{(k)} \rightarrow Q'^T AQ' = T'. \]
An easy case is for \( A \) symmetric. Recall Corollary 16.3.3. By this corollary, there exists an orthogonal (real unitary) matrix \( Q \) such that
\[ Q^T AQ = D \]
where \( D \) is diagonal having the eigenvalues on the main diagonal decreasing in size from the upper left corner to the lower right.

**Corollary 16.3.6** Let \( A \) be a real symmetric \( n \times n \) matrix having eigenvalues
\[ \lambda_1 > \lambda_2 > \cdots > \lambda_n > 0 \]
and let \( Q \) be defined by
\[ QDQ^T = A, \ D = QT AQ, \quad (16.12) \]
where \( Q \) is orthogonal and \( D \) is a diagonal matrix having the eigenvalues on the main diagonal decreasing in size from the upper left corner to the lower right. Let \( Q^T \) have an LU factorization. Then in the QR algorithm, the matrices \( Q^{(k)} \) converge to \( Q' \) where \( Q' \) is the same as \( Q \) except having some columns multiplied by \((-1)\). Thus the columns of \( Q' \) are eigenvectors of \( A \). The matrices \( A_k \) converge to \( D \).
**Proof:** This follows from Theorem 16.3.3. Here $S = Q, S^{-1} = Q^T$. Thus

$$Q = S = QR$$

and $R = I$. By Theorem 16.3.5 and Lemma 16.3.3,

$$A_k = Q^{(k)T}A Q^{(k)} \Rightarrow Q^{(k)T}A Q^{(k)} = Q^T A^k Q = D.$$  

because formula 16.12 is unaffected by replacing $Q$ with $Q'$. □

When using the QR algorithm, it is not necessary to check technical condition about $S^{-1}$ having an $LU$ factorization. The algorithm delivers a sequence of matrices which are similar to the original one. If that sequence converges to an upper triangular matrix, then the algorithm worked. Furthermore, the technical condition is sufficient but not necessary. The algorithm will work even without the technical condition.

**Example 16.3.7** Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{pmatrix} 5 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 1 \end{pmatrix}$$

It is a symmetric matrix but other than that, I just pulled it out of the air. By Lemma 16.3.3 it follows $A_k = Q^{(k)T}A Q^{(k)}$. And so to get to the answer quickly I could have the computer raise $A$ to a power and then take the QR factorization of what results to get the $k^{th}$ iteration using the above formula. Let’s pick $k = 10$.

$$\begin{pmatrix} 5 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 1 \end{pmatrix}^{10} = \begin{pmatrix} 4.2273 \times 10^7 & 2.5959 \times 10^7 & 1.8611 \times 10^7 \\ 2.5959 \times 10^7 & 1.6072 \times 10^7 & 1.1506 \times 10^7 \\ 1.8611 \times 10^7 & 1.1506 \times 10^7 & 8.2396 \times 10^6 \end{pmatrix}$$

Now take QR factorization of this. The computer will do that also.

This yields

$$\begin{pmatrix} .79785 & -.59912 & -6.6943 \times 10^{-2} \\ .48995 & .70912 & -.50706 \\ .35126 & .37176 & .85931 \end{pmatrix} \cdot \begin{pmatrix} 5.2983 \times 10^7 & 3.2627 \times 10^7 & 2.338 \times 10^7 \\ 0 & 1.2172 \times 10^5 & 71946. \\ 0 & 0 & 277.03 \end{pmatrix}$$

Next it follows

$$A_{10} = \begin{pmatrix} .79785 & -.59912 & -6.6943 \times 10^{-2} \\ .48995 & .70912 & -.50706 \\ .35126 & .37176 & .85931 \end{pmatrix}^T \cdot \begin{pmatrix} 5 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 79785 & -.59912 & -6.6943 \times 10^{-2} \\ 48995 & .70912 & -.50706 \\ 35126 & .37176 & .85931 \end{pmatrix}$$

and this equals

$$\begin{pmatrix} 6.057 & 3.698 \times 10^{-3} & 3.4346 \times 10^{-5} \\ 3.698 \times 10^{-3} & 3.2008 & -4.0643 \times 10^{-4} \\ 3.4346 \times 10^{-5} & -4.0643 \times 10^{-4} & -.2579 \end{pmatrix}$$
By Gerschgorin’s theorem, the eigenvalues are pretty close to the diagonal entries of the above matrix. Note I didn’t use the theorem, just Lemma 16.3.3 and Gerschgorin’s theorem to verify the eigenvalues are close to the above numbers. The eigenvectors are close to

\[
\begin{pmatrix}
.79785 \\
.48995 \\
.35126
\end{pmatrix}, \quad
\begin{pmatrix}
-1.59912 \\
.70912 \\
.37176
\end{pmatrix}, \quad
\begin{pmatrix}
-6.6943 \times 10^{-2} \\
-5.0706 \\
.85931
\end{pmatrix}
\]

Lets check one of these.

\[
\begin{pmatrix}
5 & 1 & 1 \\
1 & 3 & 2 \\
1 & 2 & 1
\end{pmatrix} - 6.0571 \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix} \approx
\begin{pmatrix}
.79785 \\
.48995 \\
.35126
\end{pmatrix}
\]

\[
\begin{pmatrix}
-2.1972 \times 10^{-3} \\
2.5439 \times 10^{-3} \\
1.3931 \times 10^{-3}
\end{pmatrix} \approx \begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}
\]

Now lets see how well the smallest approximate eigenvalue and eigenvector works.

\[
\begin{pmatrix}
5 & 1 & 1 \\
1 & 3 & 2 \\
1 & 2 & 1
\end{pmatrix} - (-0.2579) \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix} \approx
\begin{pmatrix}
-6.6943 \times 10^{-2} \\
-5.0706 \\
.85931
\end{pmatrix}
\]

\[
\begin{pmatrix}
2.704 \times 10^{-4} \\
-2.7377 \times 10^{-4} \\
-1.3695 \times 10^{-4}
\end{pmatrix} \approx \begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}
\]

For practical purposes, this has found the eigenvalues and eigenvectors.

### 16.3.3 The QR Algorithm In The General Case

In the case where \( A \) has distinct positive eigenvalues it was shown above that under reasonable conditions related to a certain matrix having an \( LU \) factorization the \( QR \) algorithm produces a sequence of matrices \( \{ A_k \} \) which converges to an upper triangular matrix. What if \( A \) is just an \( n \times n \) matrix having possibly complex eigenvalues but \( A \) is nondefective? What happens with the \( QR \) algorithm in this case? The short answer to this question is that the \( A_k \) of the algorithm typically cannot converge. However, this does not mean the algorithm is not useful in finding eigenvalues. It turns out the sequence of matrices \( \{ A_k \} \) have the appearance of a block upper triangular matrix for large \( k \) in the sense that the entries below the blocks on the main diagonal are small. Then looking at these blocks gives a way to approximate the eigenvalues.

First it is important to note a simple fact about unitary diagonal matrices. In what follows \( \Lambda \) will denote a unitary matrix which is also a diagonal matrix. These matrices are just the identity matrix with some of the ones replaced with a number of the form \( e^{i\theta} \) for some \( \theta \). The important property of multiplication of any matrix by \( \Lambda \) on either side is that it leaves all the zero entries the same and also preserves the absolute values of the other entries. Thus a block triangular matrix multiplied by \( \Lambda \) on either side is still block triangular. If the matrix is close to being block triangular this property of being close to a block triangular matrix is also preserved by multiplying on either side by \( \Lambda \). Other patterns depending only on the size of the absolute value occurring in the matrix are also preserved by multiplying on either side by \( \Lambda \). In other words, in looking for a pattern in a matrix, multiplication by \( \Lambda \) is irrelevant.

Now let \( A \) be an \( n \times n \) matrix having real or complex entries. By Lemma 16.3.3 and the assumption that \( A \) is nondefective, there exists an invertible \( S \),

\[
A^k = Q^{(k)} R^{(k)} = S D^k S^{-1}
\] (16.13)
where

\[
D = \begin{pmatrix}
\lambda_1 & 0 \\
\vdots & \ddots \\
0 & \cdots & \lambda_n
\end{pmatrix}
\]

and by rearranging the columns of \( S \), \( D \) can be made such that

\[|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|.
\]

Assume \( S^{-1} \) has an \( LU \) factorization. Then

\[A^k = SD^kLU = SD^kLD^{-k}D^kU.
\]

Consider the matrix in the middle, \( D^kLD^{-k} \). The \( ij^{th} \) entry is of the form

\[
(D^kLD^{-k})_{ij} = \begin{cases}
\lambda_i^k L_{ij} \lambda_j^{-k} & \text{if } j < i \\
1 & \text{if } i = j \\
0 & \text{if } j > i
\end{cases}
\]

and these all converge to 0 whenever \( |\lambda_i| < |\lambda_j| \). Thus

\[D^kLD^{-k} = (L_k + E_k)
\]

where \( L_k \) is a lower triangular matrix which has all ones down the diagonal and some subdiagonal terms of the form

\[
\lambda_i^k L_{ij} \lambda_j^{-k}
\]

for which \( |\lambda_i| = |\lambda_j| \) while \( E_k \to 0 \). (Note the entries of \( L_k \) are all bounded independent of \( k \) but some may fail to converge.) Then

\[Q^{(k)}R^{(k)} = S(L_k + E_k)D^kU
\]

Let

\[SL_k = Q_kR_k
\]

where this is the \( QR \) factorization of \( SL_k \). Then

\[
Q^{(k)}R^{(k)} = (Q_kR_k + SE_k)D^kU = Q_k(I + Q_kSE_kR_k^{-1})R_kD^kU = Q_k(I + F_k)R_kD^kU
\]

where \( F_k \to 0 \). Let \( I + F_k = Q_k' R_k' \). Then \( Q^{(k)}R^{(k)} = Q_k Q_k' R_k' R_k D^k U \). By Lemma [16.3.4]

\[Q_k' \to I \text{ and } R_k' \to I.
\]

Now let \( \Lambda_k \) be a diagonal unitary matrix which has the property that \( \Lambda_k^* D^k U \) is an upper triangular matrix which has all the diagonal entries positive. Then

\[Q^{(k)}R^{(k)} = Q_k Q_k' \Lambda_k (\Lambda_k^* R_k' R_k \Lambda_k) \Lambda_k^* D^k U
\]

That matrix in the middle has all positive diagonal entries because it is itself an upper triangular matrix, being the product of such, and is similar to the matrix \( R_k' R_k \) which is upper triangular with positive diagonal entries. By Lemma [16.3.4] again, this time using the uniqueness assertion,

\[Q^{(k)}(k) = Q_k Q_k' \Lambda_k, \quad R^{(k)} = (\Lambda_k^* R_k' R_k \Lambda_k) \Lambda_k^* D^k U
\]
Note the term $Q_k Q_k^T A_k$ must be real because the algorithm gives all $Q^{(k)}$ as real matrices. By Lemma 16.10 it follows that for $k$ large enough $Q^{(k)} \approx Q_k A_k$ where $\approx$ means the two matrices are close. Recall $A_k = Q^{(k)T} A Q^{(k)}$ and so for large $k$,

$$A_k \approx (Q_k \Lambda_k)^* A (Q_k \Lambda_k) = \Lambda_k^* Q_k^* A Q_k \Lambda_k$$

As noted above, the form of $\Lambda_k^* Q_k^* A Q_k \Lambda_k$ in terms of which entries are large and small is not affected by the presence of $\Lambda_k$ and $\Lambda_k^*$. Thus, in considering what form this is in, it suffices to consider $Q_k^* A Q_k$.

This could get pretty complicated but I will consider the case where

$$\text{if } |\lambda_i| = |\lambda_{i+1}|, \text{ then } |\lambda_{i+2}| < |\lambda_{i+1}|.$$  \hfill (16.17)

This is typical of the situation where the eigenvalues are all distinct and the matrix $A$ is real so the eigenvalues occur as conjugate pairs. Then in this case, $L_k$ above is lower triangular with some nonzero terms on the diagonal right below the main diagonal but zeros everywhere else. Thus maybe $(L_k)_{s+1,s} \neq 0$ Recall Lemma 16.15 which implies

$$Q_k = S L_k R_k^{-1}$$  \hfill (16.18)

where $R_k^{-1}$ is upper triangular. Also recall from the definition of $S$ in Lemma 16.13, it follows that $S^{-1} A S = D$. Thus the columns of $S$ are eigenvectors of $A$, the $i^{th}$ being an eigenvector for $\lambda_i$. Now from the form of $L_k$, it follows $L_k R_k^{-1}$ is a block upper triangular matrix denoted by $T_B$ and so $Q_k = S T_B$. It follows from the above construction in Lemma 16.16 and the given assumption on the sizes of the eigenvalues, there are finitely many $2 \times 2$ blocks centered on the main diagonal along with possibly some diagonal entries. Therefore, for large $k$ the matrix $A_k = Q^{(k)T} A Q^{(k)}$ is approximately of the same form as that of

$$Q_k^* A Q_k = T_B^{-1} S^{-1} A S T_B = T_B^{-1} D T_B$$

which is a block upper triangular matrix. As explained above, multiplication by the various diagonal unitary matrices does not affect this form. Therefore, for large $k$, $A_k$ is approximately a block upper triangular matrix.

How would this change if the above assumption on the size of the eigenvalues were relaxed but the matrix was still nondefective with appropriate matrices having an $LU$ factorization as above? It would mean the blocks on the diagonal would be larger. This immediately makes the problem more cumbersome to deal with. However, in the case that the eigenvalues of $A$ are distinct, the above situation really is typical of what occurs and in any case can be quickly reduced to this case.

To see this, suppose condition (16.17) is violated and $\lambda_j, \ldots, \lambda_{j+p}$ are complex eigenvalues having nonzero imaginary parts such that each has the same absolute value but they are all distinct. Then let $\mu > 0$ and consider the matrix $A + \mu I$. Thus the corresponding eigenvalues of $A + \mu I$ are $\lambda_j + \mu, \ldots, \lambda_{j+p} + \mu$. A short computation shows $|\lambda_j + \mu|, \ldots, |\lambda_{j+p} + \mu|$ are all distinct and so the above situation of (16.17) is obtained. Of course, if there are repeated eigenvalues, it may not be possible to reduce to the case above and you would end up with large blocks on the main diagonal which could be difficult to deal with.

So how do you identify the eigenvalues? You know $A_k$ and behold that it is close to a block upper triangular matrix $T_B$. You know $A_k$ is also similar to $A$. Therefore, $T_B$ has eigenvalues which are close to the eigenvalues of $A_k$ and hence those of $A$ provided $k$ is sufficiently large. See Theorem 16.22 which depends on complex analysis or the exercise on Page 346 which gives another way to see this. Thus you find the eigenvalues of this block triangular matrix $T_B$ and assert that these are good approximations of the eigenvalues of $A_k$ and hence to those of $A$. How do you find the eigenvalues of a block triangular matrix? This is easy from Lemma 16.13. Say

$$T_B = \begin{pmatrix} B_1 & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & B_m \end{pmatrix}$$
Then forming $\lambda I - T'_B$ and taking the determinant, it follows from Lemma 14.1.4 this equals

$$\prod_{j=1}^{m} \det (\lambda I_j - B_j)$$

and so all you have to do is take the union of the eigenvalues for each $B_j$. In the case emphasized here this is very easy because these blocks are just $2 \times 2$ matrices.

How do you identify approximate eigenvectors from this? First try to find the approximate eigenvectors for $A_k$. Pick an approximate eigenvalue $\lambda$, an exact eigenvalue for $T'_B$. Then find $v$ solving $T'_Bv = \lambda v$. It follows since $T'_B$ is close to $A_k$ that $A_kv \approx \lambda v$ and so

$$Q^{(k)}AQ^{(k)T}v = A_kv \approx \lambda v$$

Hence

$$AQ^{(k)T}v \approx \lambda Q^{(k)T}v$$

and so $Q^{(k)T}v$ is an approximation to the eigenvector which goes with the eigenvalue of $A$ which is close to $\lambda$.

**Example 16.3.8** Here is a matrix.

$$
\begin{pmatrix}
3 & 2 & 1 \\
-2 & 0 & -1 \\
-2 & -2 & 0
\end{pmatrix}
$$

It happens that the eigenvalues of this matrix are $1, 1 + i, 1 - i$. Lets apply the QR algorithm as if the eigenvalues were not known.

Applying the QR algorithm to this matrix yields the following sequence of matrices.

$$A_1 = \begin{pmatrix}
1.2353 & 1.9412 & 4.3657 \\
-.39215 & 1.5425 & 5.3886 \times 10^{-2} \\
-.16169 & -.18864 & .22222
\end{pmatrix}$$

$\vdots$

$$A_{12} = \begin{pmatrix}
9.1772 \times 10^{-2} & .63089 & -2.0398 \\
-2.8556 & 1.9082 & -3.1043 \\
1.0786 \times 10^{-2} & 3.4614 \times 10^{-4} & 1.0
\end{pmatrix}$$

At this point the bottom two terms on the left part of the bottom row are both very small so it appears the real eigenvalue is near 1.0. The complex eigenvalues are obtained from solving

$$\det \left( \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 9.1772 \times 10^{-2} & .63089 \\ -2.8556 & 1.9082 \end{pmatrix} \right) = 0$$

This yields

$$\lambda = 1.0 - .98828i, \ 1.0 + .98828i$$

**Example 16.3.9** The equation $x^4 + x^3 + 4x^2 + x - 2 = 0$ has exactly two real solutions. You can see this by graphing it. However, the rational root theorem from algebra shows neither of these solutions are rational. Also, graphing it does not yield any information about the complex solutions. Lets use the QR algorithm to approximate all the solutions, real and complex.
A matrix whose characteristic polynomial is the given polynomial is

\[
\begin{pmatrix}
-1 & -4 & -1 & 2 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0
\end{pmatrix}
\]

Using the QR algorithm yields the following sequence of iterates for \( A_k \)

\[
A_1 = \begin{pmatrix}
.99999 & -2.5927 & -1.7588 & -1.2978 \\
2.1213 & -1.7778 & -1.6042 & -.99415 \\
0 & .34246 & -.32749 & -.91799 \\
0 & 0 & -.44659 & .10526
\end{pmatrix}
\]

\[
\vdots
\]

\[
A_9 = \begin{pmatrix}
-.83412 & -4.1682 & -1.939 & -.7783 \\
1.05 & .14514 & .2171 & 2.5474 \times 10^{-2} \\
0 & 4.0264 \times 10^{-4} & -.85029 & -.61608 \\
0 & 0 & -1.8263 \times 10^{-2} & .53939
\end{pmatrix}
\]

Now this is similar to \( A \) and the eigenvalues are close to the eigenvalues obtained from the two blocks on the diagonal,

\[
\begin{pmatrix}
-.83412 & -4.1682 \\
1.05 & .14514
\end{pmatrix}
\begin{pmatrix}
-.85029 & -.61608 \\
-1.8263 \times 10^{-2} & .53939
\end{pmatrix}
\]

since \( 4.0264 \times 10^{-4} \) is small. After routine computations involving the quadratic formula, these are seen to be

\[-.85834, .54744, -.34449 - 2.0339i, -.34449 + 2.0339i\]

When these are plugged in to the polynomial equation, you see that each is close to being a solution of the equation.

### 16.3.4 Upper Hessenberg Matrices

It seems like most of the attention to the QR algorithm has to do with finding ways to get it to "converge" faster. Great and marvelous are the clever tricks which have been proposed to do this but my intent is to present the basic ideas, not to go in to the numerous refinements of this algorithm. However, there is one thing which should be done. It involves reducing to the case of an upper Hessenberg matrix which is one which is zero below the main sub diagonal. The following shows that any square matrix is unitarily similar to such an upper Hessenberg matrix.

Let \( A \) be an invertible \( n \times n \) matrix. Let \( Q'_1 \) be a unitary matrix

\[
Q'_1 \begin{pmatrix}
a_{21} \\
\vdots \\
a_{n1}
\end{pmatrix} = \begin{pmatrix}
\sqrt{\sum_{j=2}^{n} |a_{j1}|^2} \\
0 \\
\vdots \\
0
\end{pmatrix} = \begin{pmatrix}
a \\
0 \\
\vdots \\
0
\end{pmatrix}
\]

The vector \( Q'_1 \) is multiplying is just the bottom \( n - 1 \) entries of the first column of \( A \). Then let \( Q_1 \) be

\[
\begin{pmatrix}
1 & 0 \\
0 & Q'_1
\end{pmatrix}
\]
It follows
\[ Q_1AQ_1^* = \begin{pmatrix} 1 & 0 \\ 0 & Q_1' \end{pmatrix} AQ_1 = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & & & \vdots \\ 0 & & & A_1' \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & Q_1^* \end{pmatrix} = \begin{pmatrix} \ast & \ast & \cdots & \ast \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ddots & A_1 \\ 0 & & & \ddots \end{pmatrix} \]

Now let \( Q_2 \) be the \( n - 2 \times n - 2 \) matrix which does to the first column of \( A_1 \) the same sort of thing that the \( n - 1 \times n - 1 \) matrix \( Q_1' \) did to the first column of \( A \). Let
\[ Q_2 = \begin{pmatrix} I & 0 \\ 0 & Q_2' \end{pmatrix} \]
where \( I \) is the \( 2 \times 2 \) identity. Then applying block multiplication,
\[ Q_2Q_1AQ_1^*Q_2^* = \begin{pmatrix} \ast & \ast & \cdots & \ast \\ \ast & \ast & \cdots & \ast \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ddots & A_2 \\ 0 & & & \ddots \end{pmatrix} \]
where \( A_2 \) is now an \( n - 2 \times n - 2 \) matrix. Continuing this way you eventually get a unitary matrix \( Q \) which is a product of those discussed above such that
\[ QAQ^T = \begin{pmatrix} \ast & \ast & \cdots & \ast \\ \ast & \ast & \cdots & \ast \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ast \\ 0 & 0 & & \ast \ast \end{pmatrix} \]
This matrix equals zero below the subdiagonal. It is called an upper Hessenberg matrix.

It happens that in the QR algorithm, if \( A_k \) is upper Hessenberg, so is \( A_{k+1} \). To see this, note that the matrix is upper Hessenberg means that \( A_{ij} = 0 \) whenever \( i - j \geq 2 \).

\[ A_{k+1} = R_k Q_k \]
where \( A_k = Q_k R_k \). Therefore as shown before,
\[ A_{k+1} = R_k A_k R_k^{-1} \]
Let the \( ij^{th} \) entry of \( A_k \) be \( a_{ij}^k \). Then if \( i - j \geq 2 \)
\[ a_{ij}^{k+1} = \sum_{p=i}^{n} \sum_{q=i}^{j} r_{ip} a_{pq}^k r_{qj}^{-1} \]
It is given that \( a_{pq}^k = 0 \) whenever \( p - q \geq 2 \). However, from the above sum,
\[ p - q \geq i - j \geq 2 \]
and so the sum equals 0.

Since upper Hessenberg matrices stay that way in the algorithm and it is closer to being upper triangular, it is reasonable to suppose the QR algorithm will yield good results more quickly for this upper Hessenberg matrix than for the original matrix. This would be especially true if the matrix is good sized. The other important thing to observe is that, starting with an upper Hessenberg matrix, the algorithm will restrict the size of the blocks which occur to being $2 \times 2$ blocks which are easy to deal with. These blocks allow you to identify the eigenvalues.

**Example 16.3.10** Let $A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & -2 & -3 & 3 \\ -3 & 5 & 1 \end{pmatrix}$ a symmetric matrix. Thus it has real eigenvalues and can be diagonalized. Find its eigenvalues.

As explained above, there is an upper Hessenberg matrix. Matlab can find it using the techniques given above pretty quickly. The syntax is as follows.

```matlab
A=[2 1 3;-5,3,-2;1,2,3];
[P,H]=hess(A)
```

Then the Hessenberg matrix similar to $A$ is

$$H = \begin{pmatrix} -1.4476 & -4.9048 & 0 & 0 \\ -4.9048 & 3.2553 & -2.0479 & 0 \\ 0 & -2.0479 & 2.1923 & -5.0990 \\ 0 & 0 & -5.0990 & -3 \end{pmatrix}$$

Note how it is symmetric also. This will always happen when you begin with a symmetric matrix. Now use the QR algorithm on this matrix. The syntax is as follows in Matlab.

```matlab
H=[enter H here]
hold on
for k=1:100
[Q,R]=qr(H);
H=R*Q;
end
Q
R
H
```

You already have $H$ and matlab knows about it so you don’t need to enter $H$ again. This yields the following matrix similar to the original one.

$$\begin{pmatrix} 7.4618 & 0 & 0 & 0 \\ 0 & -6.3804 & 0 & 0 \\ 0 & 0 & -4.419 & -3.679 \\ 0 & 0 & -3.679 & 4.3376 \end{pmatrix}$$

The eigenvalues of this matrix are

$$7.4618, -6.3804, 4.353, -4.4344$$

You might want to check that the product of these equals the determinant of the matrix and that the sum equals the trace of the matrix. In fact, this works out very well. To find eigenvectors, you could use the shifted inverse power method.
CHAPTER 16. NUMERICAL METHODS, EIGENVALUES

16.4 Exercises

In these exercises which call for a computation, don’t waste time on them unless you use a computer or calculator which can raise matrices to powers and take QR factorizations.

1. In Example 16.2.7 an eigenvalue was found correct to several decimal places along with an eigenvector. Find the other eigenvalues along with their eigenvectors.

2. Find the eigenvalues and eigenvectors of the matrix \( A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 1 & 3 \\ 1 & 3 & 2 \end{pmatrix} \) numerically. In this case the exact eigenvalues are \( \pm \sqrt{3}, 6 \). Compare with the exact answers.

3. Find the eigenvalues and eigenvectors of the matrix \( A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 2 \end{pmatrix} \) numerically. The exact eigenvalues are \( 2, 4 + \sqrt{15}, 4 - \sqrt{15} \). Compare your numerical results with the exact values. Is it much fun to compute the exact eigenvectors?

4. Find the eigenvalues and eigenvectors of the matrix \( A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 3 & 3 \\ 1 & 3 & 2 \end{pmatrix} \) numerically. I don’t know the exact eigenvalues in this case. Check your answers by multiplying your numerically computed eigenvectors by the matrix.

5. Find the eigenvalues and eigenvectors of the matrix \( A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 0 & 3 \\ 1 & 3 & 2 \end{pmatrix} \) numerically. I don’t know the exact eigenvalues in this case. Check your answers by multiplying your numerically computed eigenvectors by the matrix.

6. Consider the matrix \( A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 0 \end{pmatrix} \) and the vector \((1, 1, 1)^T\). Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of \( A \).

7. Consider the matrix \( A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 1 & 4 \\ 1 & 4 & 5 \end{pmatrix} \) and the vector \((1, 1, 1)^T\). Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of \( A \).

8. Consider the matrix \( A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 6 & 4 \\ 3 & 4 & -3 \end{pmatrix} \) and the vector \((1, 1, 1)^T\). Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of \( A \).

9. Using Gerschgorin’s theorem, find upper and lower bounds for the eigenvalues of \( A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 6 & 4 \\ 3 & 4 & -3 \end{pmatrix} \).

10. Tell how to find a matrix whose characteristic polynomial is a given monic polynomial. This is called a companion matrix. Find the roots of the polynomial \( x^3 + 7x^2 + 3x + 7 \).

11. Find the roots to \( x^4 + 3x^3 + 4x^2 + x + 1 \). It has two complex roots.
12. Suppose $A$ is a real symmetric matrix and the technique of reducing to an upper Hessenberg matrix is followed. Show the resulting upper Hessenberg matrix is actually equal to 0 on the top as well as the bottom.
Part III

Analysis Which Involves Linear Algebra
Chapter 17

The Derivative, A Linear Transformation

In this part of the book, some topics in analysis are considered in which linear algebra plays a key role. This is not about linear algebra, but linear algebra is used in a very essential manner.

17.1 Basic Definitions

The derivative is a linear transformation. This may not be entirely clear from a beginning calculus course because they like to say it is a slope which is a number. As observed by Deudonné, “...In the classical teaching of Calculus, this idea (that the derivative is a linear transformation) is immediately obscured by the accidental fact that, on a one-dimensional vector space, there is a one-to-one correspondence between linear forms and numbers, and therefore the derivative at a point is defined as a number instead of a linear form. This slavish subservience to the shibboleth of numerical interpretation at any cost becomes much worse when dealing with functions of several variables...” He is absolutely right.

The concept of derivative generalizes right away to functions of many variables but only if you regard a number which is identified as the derivative in single variable calculus as a linear transformation on \( \mathbb{R} \). However, no attempt will be made to consider derivatives from one side or another. This is because when you consider functions of many variables, there isn’t a well defined side. However, it is certainly the case that there are more general notions which include such things. I will present a fairly general notion of the derivative of a function which is defined on a normed vector space which has values in a normed vector space. The case of most interest is that of a function which maps an open set in \( \mathbb{F}^n \) to \( \mathbb{F}^m \) but it is no more trouble to consider the extra generality and it is sometimes useful to have this extra generality because sometimes you want to consider functions defined, for example on subspaces of \( \mathbb{F}^n \) and it is nice to not have to trouble with ad hoc considerations. Also, you might want to consider \( \mathbb{F}^n \) with some norm other than the usual one.

For most of what follows, it is not important for the vector spaces to be finite dimensional provided you make the following definition which is automatic in finite dimensions.

**Definition 17.1.1** Let \((X, \|\cdot\|_X)\) and \((Y, \|\cdot\|_Y)\) be two normed linear spaces. Then \( \mathcal{L}(X, Y) \) denotes the set of linear maps from \( X \) to \( Y \) which also satisfy the following condition. For \( L \in \mathcal{L}(X, Y) \),

\[
\lim_{\|x\|_X \leq 1} \|Lx\|_Y \equiv \|L\| < \infty
\]

---

405

1In the Bible, there was a battle between Ephraimites and Gilleadites during the time of Jephthah, the judge who sacrificed his daughter to Jehovah, one of several instances of human sacrifice in the Bible. The cause of this battle was very strange. However, the Ephraimites lost and when they tried to cross a river to get back home, they had to say shibboleth. If they said “sibboleth” they were killed because their inability to pronounce the “sh” sound identified them as Ephraimites. They usually don’t tell this story in Sunday school. The word has come to denote something which is arbitrary and no longer important.
To save notation, I will use $\| \cdot \|$ as a norm on either $X$ or $Y$ and allow the context to determine which it is.

Let $U$ be an open set in $X$, and let $f : U \to Y$ be a function.

**Definition 17.1.2** A function $g$ is $o(v)$ if

$$
\lim_{\|v\| \to 0} \frac{g(v)}{\|v\|} = 0
$$

(17.1)

A function $f : U \to Y$ is differentiable at $x \in U$ if there exists a linear transformation $L \in \mathcal{L}(X,Y)$ such that

$$
f(x + v) = f(x) + Lv + o(v)
$$

This linear transformation $L$ is the definition of $Df(x)$. This derivative is often called the Frechet derivative.

Note that from Theorem [17.1.3] the question whether a given function is differentiable is independent of the norm used on the finite dimensional vector space. That is, it is a function that is differentiable with one norm if and only if it is differentiable with another norm. In infinite dimensions, this is not clearly so and in this case, simply regard the norm as part of the definition of the normed linear space which incidentally will also typically be assumed to be a complete normed linear space.

The definition [17.1.3] means the error,

$$
f(x + v) - f(x) - Lv
$$

converges to 0 faster than $\|v\|$. Thus the above definition is equivalent to saying

$$
\lim_{\|v\| \to 0} \frac{\|f(x + v) - f(x) - Lv\|}{\|v\|} = 0
$$

(17.2)

or equivalently,

$$
\lim_{y \to x} \frac{\|f(y) - f(x) - Df(x)(y - x)\|}{\|y - x\|} = 0.
$$

(17.3)

The symbol, $o(v)$ should be thought of as an adjective. Thus, if $t$ and $k$ are constants,

$$
o(v) = o(v) + o(v), \quad o(tv) = o(v), \quad ko(v) = o(v)
$$

and other similar observations hold.

**Theorem 17.1.3** The derivative is well defined.

**Proof:** First note that for a fixed vector, $v$, $o(tv) = o(t)$. This is because

$$
\lim_{t \to 0} \frac{o(tv)}{|t|} = \lim_{t \to 0} \frac{\|v\| o(tv)}{\|tv\|} = 0
$$

Now suppose both $L_1$ and $L_2$ work in the above definition. Then let $v$ be any vector and let $t$ be a real scalar which is chosen small enough that $tv + x \in U$. Then

$$
f(x + tv) = f(x) + L_1 tv + o(tv), \quad f(x + tv) = f(x) + L_2 tv + o(tv).
$$

Therefore, subtracting these two yields $(L_2 - L_1)(tv) = o(tv) = o(t)$. Therefore, dividing by $t$ yields $(L_2 - L_1)(v) = \frac{o(t)}{t}$. Now let $t \to 0$ to conclude that $(L_2 - L_1)(v) = 0$. Since this is true for all $v$, it follows $L_2 = L_1$. This proves the theorem.

**Lemma 17.1.4** Let $f$ be differentiable at $x$. Then $f$ is continuous at $x$ and in fact, there exists $K > 0$ such that whenever $\|v\|$ is small enough,

$$
\|f(x + v) - f(x)\| \leq K \|v\|
$$

Also if $f$ is differentiable at $x$, then

$$
o(\|f(x + v) - f(x)\|) = o(v)
$$


**Proof:** From the definition of the derivative,

\[ f(x + v) - f(x) = Df(x)v + o(v). \]

Let \( ||v|| \) be small enough that \( \frac{o(||v||)}{||v||} < 1 \) so that \( ||o(v)|| \leq ||v|| \). Then for such \( v \),

\[
||f(x + v) - f(x)|| \leq ||Df(x)v|| + ||v||
\leq (||Df(x)|| + 1) ||v||
\]

This proves the lemma with \( K = ||Df(x)|| + 1 \). Recall the operator norm discussed in Definitions 17.1.1 17.1.1.

The last assertion is implied by the first as follows. Define

\[
h(v) = \begin{cases}
    \frac{o(||f(x + v) - f(x)||)}{||f(x + v) - f(x)||} & \text{if } ||f(x + v) - f(x)|| \neq 0 \\
    0 & \text{if } ||f(x + v) - f(x)|| = 0
\end{cases}
\]

Then \( \lim_{||v|| \to 0} h(v) = 0 \) from continuity of \( f \) at \( x \) which is implied by the first part. Also from the above estimate,

\[
\frac{o(||f(x + v) - f(x)||)}{||v||} = ||h(v)|| \frac{||f(x + v) - f(x)||}{||v||} \leq ||h(v)|| (||Df(x)|| + 1)
\]

This establishes the second claim. ■

Here \( ||Df(x)|| \) is the operator norm of the linear transformation, \( Df(x) \).

## 17.2 The Chain Rule

With the above lemma, it is easy to prove the chain rule.

**Theorem 17.2.1 (The chain rule)** Let \( U \) and \( V \) be open sets \( U \subseteq X \) and \( V \subseteq Y \). Suppose \( f : U \to V \) is differentiable at \( x \in U \) and suppose \( g : V \to \mathbb{F}^q \) is differentiable at \( f(x) \in V \). Then \( g \circ f \) is differentiable at \( x \) and

\[
D(g \circ f)(x) = Dg(f(x))Df(x).
\]

**Proof:** This follows from a computation. Let \( B(x,r) \subseteq U \) and let \( r \) also be small enough that for \( ||v|| \leq r \), it follows that \( f(x + v) \in V \). Such an \( r \) exists because \( f \) is continuous at \( x \). For \( ||v|| < r \), the definition of differentiability of \( g \) and \( f \) implies

\[
g(f(x + v)) - g(f(x)) =
\]

\[
Dg(f(x))(f(x + v) - f(x)) + o(f(x + v) - f(x))
\]

\[
= Dg(f(x))Df(x)v + o(v)
\]

\[
= D(g(f(x)))D(f(x))v + o(v)
\]

\[
= D(g(f(x)))D(f(x))v + o(v)
\]

By Lemma 17.1.1. From the definition of the derivative \( D(g \circ f)(x) \) exists and equals \( D(g(f(x)))D(f(x)) \).

### ■

## 17.3 The Matrix Of The Derivative

The case of interest here is where \( X = \mathbb{R}^n \) and \( Y = \mathbb{R}^m \), the function being defined on an open subset of \( \mathbb{R}^n \). Of course this all generalizes to arbitrary vector spaces and one considers the matrix
CHAPTER 17. THE DERIVATIVE, A LINEAR TRANSFORMATION

taken with respect to various bases. As above, \( f \) will be defined and differentiable on an open set \( U \subseteq \mathbb{R}^n \).

As discussed in the review material on linear maps, the matrix of \( Df(x) \) is the matrix having the \( i^{th} \) column equal to \( Df(x)e_i \) and so it is only necessary to compute this. Let \( t \) be a small real number such that both

\[
\frac{f(x + te_i) - f(x) - Df(x)(te_i)}{t} = o(t)
\]

Therefore,

\[
\frac{f(x + te_i) - f(x)}{t} = Df(x)(e_i) + o(t)
\]

The limit exists on the right and so it exists on the left also. Thus

\[
\frac{\partial f(x)}{\partial x_i} = \lim_{t \to 0} \frac{f(x + te_i) - f(x)}{t} = Df(x)(e_i)
\]

and so the matrix of the derivative is just the matrix which has the \( i^{th} \) column equal to the \( i^{th} \) partial derivative of \( f \). Note that this shows that whenever \( f \) is differentiable, it follows that the partial derivatives all exist. It does not go the other way however as discussed later.

**Theorem 17.3.1** Let \( f : U \subseteq \mathbb{F}^n \to \mathbb{F}^m \) and suppose \( f \) is differentiable at \( x \). Then all the partial derivatives \( \frac{\partial f_i(x)}{\partial x_j} \) exist and if \( Jf(x) \) is the matrix of the linear transformation, \( Df(x) \) with respect to the standard basis vectors, then the \( ij^{th} \) entry is given by \( \frac{\partial f_i}{\partial x_j}(x) \) also denoted as \( f_{i,j} \) or \( f_i;x_j \). It is the matrix whose \( i^{th} \) column is

\[
\frac{\partial f_i(x)}{\partial x_j} = \lim_{t \to 0} \frac{f(x + te_i) - f(x)}{t} = Df(x)(e_i)
\]

Of course there is a generalization of this idea called the directional derivative.

**Definition 17.3.2** In general, the symbol

\[
D_vf(x)
\]

is defined by

\[
\lim_{t \to 0} \frac{f(x + tv) - f(x)}{t}
\]

where \( t \in \mathbb{F} \). In case \(|v| = 1\) and the norm is the standard Euclidean norm, this is called the directional derivative. More generally, with no restriction on the size of \( v \) and in any linear space, it is called the Gateaux derivative. \( f \) is said to be Gateaux differentiable at \( x \) if there exists \( D_vf(x) \) such that

\[
\lim_{t \to 0} \frac{f(x + tv) - f(x)}{t} = D_vf(x)
\]

where \( v \to D_vf(x) \) is linear. Thus we say it is Gateaux differentiable if the Gateaux derivative exists for each \( v \) and \( v \to D_vf(x) \) is linear.

What if all the partial derivatives of \( f \) exist? Does it follow that \( f \) is differentiable? Consider the following function, \( f : \mathbb{R}^2 \to \mathbb{R} \),

\[
f(x, y) = \begin{cases} \frac{xy}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}
\]

\(^2\)René Gateaux was one of the many young French men killed in world war I. This derivative is named after him, but it developed naturally from ideas used in the calculus of variations which were due to Euler and Lagrange back in the 1700's.
Then from the definition of partial derivatives,
\[ \lim_{h \to 0} \frac{f(h, 0) - f(0, 0)}{h} = \lim_{h \to 0} \frac{0 - 0}{h} = 0 \]
and
\[ \lim_{h \to 0} \frac{f(0, h) - f(0, 0)}{h} = \lim_{h \to 0} \frac{0 - 0}{h} = 0 \]
However, \( f \) is not even continuous at \((0, 0)\) which may be seen by considering the behavior of the function along the line \( y = x \) and along the line \( x = 0 \). By Lemma 17.1.4 this implies \( f \) is not differentiable. Therefore, it is necessary to consider the correct definition of the derivative given above if you want to get a notion which generalizes the concept of the derivative of a function of one variable in such a way as to preserve continuity whenever the function is differentiable.

17.4 A Mean Value Inequality

The following theorem will be very useful in much of what follows. It is a version of the mean value theorem as is the next lemma.

**Lemma 17.4.1** Let \( Y \) be a normed vector space and suppose \( h : [0, 1] \to Y \) is differentiable and satisfies
\[ \|h'(t)\| \leq M. \]
Then
\[ \|h(1) - h(0)\| \leq M. \]
**Proof:** Let \( \varepsilon > 0 \) be given and let
\[ S = \{ t \in [0,1] : \text{ for all } s \in [0, t], \|h(s) - h(0)\| \leq (M + \varepsilon) s \} \]
Then \( 0 \in S \). Let \( t = \sup S \). Then by continuity of \( h \) it follows
\[ \|h(t) - h(0)\| = (M + \varepsilon) t \quad (17.5) \]
Suppose \( t < 1 \). Then there exist positive numbers, \( h_k \) decreasing to 0 such that
\[ \|h(t + h_k) - h(0)\| > (M + \varepsilon) (t + h_k) \]
and now it follows from (17.5) and the triangle inequality that
\[ \|h(t + h_k) - h(t)\| + \|h(t) - h(0)\| \]
\[ = \|h(t + h_k) - h(t)\| + (M + \varepsilon) t > (M + \varepsilon) (t + h_k) \]
and so
\[ \|h(t + h_k) - h(t)\| > (M + \varepsilon) h_k \]
Now dividing by \( h_k \) and letting \( k \to \infty \)
\[ \|h'(t)\| \geq M + \varepsilon, \]
a contradiction. Thus \( t = 1 \). \( \blacksquare \)

**Theorem 17.4.2** Suppose \( U \) is an open subset of \( X \) and \( f : U \to Y \) has the property that \( Df(x) \) exists for all \( x \in U \) and that, \( x + t(y - x) \in U \) for all \( t \in [0,1] \). (The line segment joining the two points lies in \( U \).) Suppose also that for all points on this line segment,
\[ \|Df(x+t(y-x))\| \leq M. \]
Then
\[ \|f(y) - f(x)\| \leq M \|y - x\|. \]
**Proof:** Let \( h(t) \equiv f(x + t(y - x)). \)

Then by the chain rule,
\[
h'(t) = Df(x + t(y - x)) (y - x)
\]
and so
\[
||h'(t)|| = ||Df(x + t(y - x))(y - x)||
\leq M||y - x||
\]
by Lemma [7.4.1]
\[
||h(1) - h(0)|| = ||f(y) - f(x)|| \leq M||y - x||. \]

### 17.5 Existence Of The Derivative, \( C^1 \) Functions

There is a way to get the differentiability of a function from the existence and continuity of the Gateaux derivatives. This is very convenient because these Gateaux derivatives are taken with respect to a one dimensional variable. The following theorem is the main result.

**Theorem 17.5.1** Let \( X \) be a normed vector space having basis \( \{v_1, \cdots, v_n\} \) and let \( Y \) be another normed vector space having basis \( \{w_1, \cdots, w_m\} \). Let \( U \) be an open set in \( X \) and let \( f: U \to Y \) have the property that the Gateaux derivatives,
\[
D_{v_k} f(x) \equiv \lim_{t \to 0} \frac{f(x + tv_k) - f(x)}{t}
\]
exist and are continuous functions of \( x \). Then \( Df(x) \) exists and
\[
Df(x) v = \sum_{k=1}^{n} D_{v_k} f(x) a_k
\]
where
\[
v = \sum_{k=1}^{n} a_k v_k.
\]
Furthermore, \( x \to Df(x) \) is continuous; that is
\[
\lim_{y \to x} ||Df(y) - Df(x)|| = 0.
\]

**Proof:** Let \( v = \sum_{k=1}^{n} a_k v_k \). Then
\[
f(x + v) - f(x) = f\left(x + \sum_{k=1}^{n} a_k v_k\right) - f(x).
\]

Then letting \( \sum_{k=1}^{0} = 0 \), \( f(x + v) - f(x) \) is given by
\[
\sum_{k=1}^{n} \left[ f\left(x + \sum_{j=1}^{k} a_j v_j\right) - f\left(x + \sum_{j=1}^{k-1} a_j v_j\right) \right]
\]
\[
= \sum_{k=1}^{n} [f(x + a_k v_k) - f(x)] + \sum_{k=1}^{n} \left[ f\left(x + \sum_{j=1}^{k} a_j v_j\right) - f(x + a_k v_k) \right] - \left[ f\left(x + \sum_{j=1}^{k-1} a_j v_j\right) - f(x) \right]
\]
(17.6)
Consider the $k^{th}$ term in (17.7). Let

$$h(t) = f \left( x + \sum_{j=1}^{k-1} a_j v_j + t a_k v_k \right) - f (x + t a_k v_k)$$

for $t \in [0, 1]$. Then

$$h' (t) = a_k \lim_{h \to 0} \frac{1}{a_k h} \left( f \left( x + \sum_{j=1}^{k-1} a_j v_j + (t + h) a_k v_k \right) - f (x + (t + h) a_k v_k) \right)$$

$$- \left( f \left( x + \sum_{j=1}^{k-1} a_j v_j + t a_k v_k \right) - f (x + t a_k v_k) \right)$$

and this equals

$$\left( D_{v_k} f \left( x + \sum_{j=1}^{k-1} a_j v_j + t a_k v_k \right) - D_{v_k} f (x + t a_k v_k) \right) a_k$$

(17.7)

Now without loss of generality, it can be assumed that the norm on $X$ is given by

$$||v|| = \max \left\{ |a_k| : v = \sum_{j=1}^{n} a_k v_k \right\}$$

because by Theorem 17.5.3, all norms on $X$ are equivalent. Therefore, from (17.7) and the assumption that the Gateaux derivatives are continuous,

$$||h' (t)|| = \left\| \left( D_{v_k} f \left( x + \sum_{j=1}^{k-1} a_j v_j + t a_k v_k \right) - D_{v_k} f (x + t a_k v_k) \right) a_k \right\|$$

$$\leq \varepsilon |a_k| \leq \varepsilon ||v||$$

provided $||v||$ is sufficiently small. Since $\varepsilon$ is arbitrary, it follows from Lemma 17.5.2 that the expression in (17.7) is $o(\varepsilon)$ because this expression equals a finite sum of terms of the form $h(1) - h(0)$ where $||h'(t)|| \leq \varepsilon ||v||$ whenever $||v||$ is small enough. Thus

$$f (x + v) - f (x) = \sum_{k=1}^{n} [f (x + a_k v_k) - f (x)] + o(\varepsilon)$$

$$= \sum_{k=1}^{n} D_{v_k} f(x) a_k + \sum_{k=1}^{n} [f (x + a_k v_k) - f (x) - D_{v_k} f (x) a_k] + o(\varepsilon).$$

Consider the $k^{th}$ term in the second sum.

$$f (x + a_k v_k) - f (x) - D_{v_k} f (x) a_k = a_k \left( \frac{f (x + a_k v_k) - f (x)}{a_k} - D_{v_k} f (x) \right)$$

where the expression in the parentheses converges to 0 as $a_k \to 0$. Thus whenever $||v||$ is sufficiently small,

$$||f (x + a_k v_k) - f (x) - D_{v_k} f (x) a_k|| \leq \varepsilon |a_k| \leq \varepsilon ||v||$$

which shows the second sum is also $o(\varepsilon)$. Therefore,

$$f (x + v) - f (x) = \sum_{k=1}^{n} D_{v_k} f(x) a_k + o(\varepsilon).$$
Defining

\[ Df(x) v = \sum_{k=1}^{n} Dv_k f(x) a_k \]

where \( v = \sum_k a_k v_k \), it follows \( Df(x) \in \mathcal{L}(X, Y) \) and is given by the above formula.

It remains to verify \( x \to Df(x) \) is continuous.

\[
\left\| (Df(x) - Df(y)) v \right\| \\
\leq \sum_{k=1}^{n} \left\| (Dv_k f(x) - Dv_k f(y)) a_k \right\| \\
\leq \max \left\{ |a_k|, k = 1, \ldots, n \right\} \sum_{k=1}^{n} \left\| Dv_k f(x) - Dv_k f(y) \right\| \\
= \|v\| \sum_{k=1}^{n} \left\| Dv_k f(x) - Dv_k f(y) \right\|
\]

and so

\[
\left\| Df(x) - Df(y) \right\| \leq \sum_{k=1}^{n} \left\| Dv_k f(x) - Dv_k f(y) \right\|
\]

which proves the continuity of \( Df \) because of the assumption the Gateaux derivatives are continuous.

This motivates the following definition of what it means for a function to be \( C^1 \).

**Definition 17.5.2** Let \( U \) be an open subset of a normed finite dimensional vector space, \( X \) and let \( f: U \to Y \) another finite dimensional normed vector space. Then \( f \) is said to be \( C^1 \) if there exists a basis for \( X, \{v_1, \ldots, v_n\} \) such that the Gateaux derivatives,

\[ Dv_k f(x) \]

exist on \( U \) and are continuous.

Note that as a special case where \( X = \mathbb{R}^n \), you could let the \( v_k = e_k \) and the condition would reduce to nothing more than a statement that the partial derivatives \( \frac{\partial f}{\partial x_i} \) are all continuous.

Here is another definition of what it means for a function to be \( C^1 \).

**Definition 17.5.3** Let \( U \) be an open subset of a normed vector space, \( X \) and let \( f: U \to Y \) another normed vector space. Then \( f \) is said to be \( C^1 \) if \( f \) is differentiable and \( x \to Df(x) \) is continuous as a map from \( U \) to \( \mathcal{L}(X, Y) \).

Now the following major theorem states these two definitions are equivalent. This is obviously so in the special case where \( X = \mathbb{R}^n \) and the special basis is the usual one because, as observed earlier, the matrix of \( Df(x) \) is just the one which has for its columns the partial derivatives which are given to be continuous.

**Theorem 17.5.4** Let \( U \) be an open subset of a normed finite dimensional vector space, \( X \) and let \( f: U \to Y \) another finite dimensional normed vector space. Then the two definitions above are equivalent.

**Proof:** It was shown in Theorem 17.5.1, the one about the continuity of the Gateaux derivatives yielding differentiability, that Definition 17.5.2 implies 17.5.3. Suppose then that Definition 17.5.3 holds. Then if \( v \) is any vector,

\[
\lim_{t \to 0} \frac{f(x + tv) - f(x)}{t} = \lim_{t \to 0} \frac{Df(x) tv + o(tv)}{t} = Df(x) v + \lim_{t \to 0} \frac{o(tv)}{t} = Df(x) v
\]
Thus $D_v f(x)$ exists and equals $Df(x)v$. By continuity of $x \to Df(x)$, this establishes continuity of $x \to D_v f(x)$ and proves the theorem. ■

Note that the proof of the theorem also implies the following corollary.

**Corollary 17.5.5** Let $U$ be an open subset of a normed finite dimensional vector space, $X$ and let $f : U \to Y$ another finite dimensional normed vector space. Then if there is a basis of $X, \{v_1, \ldots, v_n\}$ such that the Gateaux derivatives, $D_{v_k} f(x)$ exist and are continuous. Then all Gateaux derivatives, $D_v f(x)$ exist and are continuous for all $v \in X$.

From now on, whichever definition is more convenient will be used.

## 17.6 Higher Order Derivatives

If $f : U \subseteq X \to Y$ for $U$ an open set, then

$$x \to Df(x)$$

is a mapping from $U$ to $\mathcal{L}(X,Y)$, a normed vector space. Therefore, it makes perfect sense to ask whether this function is also differentiable.

**Definition 17.6.1** The following is the definition of the second derivative.

$$D^2 f(x) \equiv D(Df(x)).$$

Thus,

$$Df(x + v) - Df(x) = D^2 f(x)v + o(v).$$

This implies

$$D^2 f(x) \in \mathcal{L}(X, \mathcal{L}(X,Y)), \ D^2 f(x)(u)(v) \in Y,$$

and the map

$$(u, v) \to D^2 f(x)(u)(v)$$

is a bilinear map having values in $Y$. In other words, the two functions,

$$u \to D^2 f(x)(u)(v), \ v \to D^2 f(x)(u)(v)$$

are both linear.

The same pattern applies to taking higher order derivatives. Thus,

$$D^3 f(x) \equiv D(D^2 f(x))$$

and $D^3 f(x)$ may be considered as a trilinear map having values in $Y$. In general $D^k f(x)$ may be considered a $k$ linear map. This means the function

$$(u_1, \ldots, u_k) \to D^k f(x)(u_1) \cdots (u_k)$$

has the property

$$u_j \to D^k f(x)(u_1) \cdots (u_j) \cdots (u_k)$$

is linear.

Also, instead of writing

$$D^2 f(x)(u)(v), \ or \ D^3 f(x)(u)(v)(w)$$

the following notation is often used.

$$D^2 f(x)(u, v) \ or \ D^3 f(x)(u, v, w)$$
with similar conventions for higher derivatives than 3. Another convention which is often used is the notation
\[ D^k f(x) v^k \]
instead of
\[ D^k f(x)(v, \ldots, v). \]

Note that for every \( k \), \( D^k f \) maps \( U \) to a normed vector space. As mentioned above, \( Df(x) \) has values in \( \mathcal{L}(X,Y) \), \( D^2 f(x) \) has values in \( \mathcal{L}(X, \mathcal{L}(X,Y)) \), etc. Thus it makes sense to consider whether \( D^k f \) is continuous. This is described in the following definition.

**Definition 17.6.2** Let \( U \) be an open subset of \( X \), a normed vector space, and let \( f: U \to Y \). Then \( f \) is \( C^k(U) \) if \( f \) and its first \( k \) derivatives are all continuous. Also, \( D^k f(x) \) when it exists can be considered a \( Y \) valued multi-linear function. Sometimes these are called tensors in case \( f \) has scalar values.

### 17.7 Some Standard Notation

In the case where \( X = \mathbb{R}^n \) there is a special notation which is often used to describe higher order mixed partial derivatives. It is called multi-index notation.

**Definition 17.7.1** \( \alpha = (\alpha_1, \ldots, \alpha_n) \) for \( \alpha_1 \cdots \alpha_n \) positive integers is called a multi-index. For \( \alpha \) a multi-index, \( |\alpha| = \alpha_1 + \cdots + \alpha_n \), and if \( x \in X \),
\[
    x = (x_1, \ldots, x_n),
\]
and \( f \) a function, define
\[
x^{\alpha} = x_1^{\alpha_1}x_2^{\alpha_2} \cdots x_n^{\alpha_n}, \quad D^{\alpha} f(x) = \frac{\partial^{|\alpha|} f(x)}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}}.
\]

Then in this special case, the following is another description of what is meant by a \( C^k \) function.

**Definition 17.7.2** Let \( U \) be an open subset of \( \mathbb{R}^n \) and let \( f: U \to Y \). Then for \( k \) a nonnegative integer, a differentiable function \( f \) is \( C^k \) if for every \( |\alpha| \leq k \), \( D^{\alpha} f \) exists and is continuous.

**Theorem 17.7.3** Let \( U \) be an open subset of \( \mathbb{R}^n \) and let \( f: U \to Y \). Then if \( D^r f(x) \) exists for \( r \leq k \), then \( D^r f \) is continuous at \( x \) for \( r \leq k \) if and only if \( D^{\alpha} f \) is continuous at \( x \) for each \( |\alpha| \leq k \).

**Proof:** First consider the case of a single derivative. Then as shown above, the matrix of \( Df(x) \) is just
\[
    J(x) = \left( \begin{array}{c} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{array} \right)
\]
and to say that \( x \to Df(x) \) is continuous is the same as saying that each of these partial derivatives is continuous. Written out in more detail,
\[
    f(x + v) - f(x) = Df(x)v + o(v) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x)v_k + o(v)
\]
Thus
\[
    Df(x)v = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x)v_k
\]
Now go to the second derivative.
\[
    D^2f(x)(w)(v) = \sum_{k=1}^{n} \left( \frac{\partial^2 f}{\partial x_k^2}(x + w) - \frac{\partial f}{\partial x_k}(x) \right)v_k + o(w)(v)
\]
\[
= \sum_{k=1}^{n} \left( \sum_{j=1}^{n} \frac{\partial^2 f(x)}{\partial x_j \partial x_k} w_j + o(w) \right) v_k + o(w)(v)
\]

and so
\[
D^2 f(x)(v) = \sum_{j,k} \frac{\partial^2 f(x)}{\partial x_j \partial x_k} w_j v_k
\]

Hence \(D^2 f\) is continuous if and only if each of these coefficients
\[
x \mapsto \frac{\partial^2 f(x)}{\partial x_j \partial x_k}
\]
is continuous. Obviously you can continue doing this and conclude that \(D^k f\) is continuous if and only if all of the partial derivatives of order up to \(k\) are continuous. \(\blacksquare\)

In practice, this is usually what people are thinking when they say that \(f\) is \(C^k\). But as just argued, this is the same as saying that the \(r\) linear form \(x \mapsto D^r f(x)\) is continuous into the appropriate space of linear transformations for each \(r \leq k\).

Of course the above is based on the assumption that the first \(k\) derivatives exist and gives two equivalent formulations which state that these derivatives are continuous. Can anything be said about the existence of the derivatives based on the existence and continuity of the partial derivatives? This is in the next section.

17.8 The Derivative And The Cartesian Product

There are theorems which can be used to get differentiability of a function based on existence and continuity of the partial derivatives. A generalization of this was given above. Here a function defined on a product space is considered. It is very much like what was presented above and could be obtained as a special case but to reinforce the ideas, I will do it from scratch because certain aspects of it are important in the statement of the implicit function theorem.

The following is an important abstract generalization of the concept of partial derivative presented above. Instead of taking the derivative with respect to one variable, it is taken with respect to several but not with respect to others. This vague notion is made precise in the following definition. First here is a lemma.

**Lemma 17.8.1** Suppose \(U\) is an open set in \(X \times Y\). Then the set, \(U_y\) defined by
\[
U_y = \{ x \in X : (x, y) \in U \}
\]
is an open set in \(X\). Here \(X \times Y\) is a finite dimensional vector space in which the vector space operations are defined componentwise. Thus for \(a, b \in \mathbb{F}\),
\[
a(x_1, y_1) + b(x_2, y_2) = (ax_1 + bx_2, ay_1 + by_2)
\]
and the norm can be taken to be
\[
\|(x, y)\| = \max (\|x\|, \|y\|)
\]

**Proof:** Recall by Theorem 11.5.3 it does not matter how this norm is defined and the definition above is convenient. It obviously satisfies most axioms of a norm. The only one which is not obvious is the triangle inequality. I will show this now.

\[
\|(x, y) + (x_1, y_1)\| \equiv \|(x + x_1, y + y_1)\| \equiv \max (\|x + x_1\|, \|y + y_1\|) \\
\leq \max (\|x\| + \|x_1\|, \|y\| + \|y_1\|) \\
\leq \max (\|x\|, \|y\|) + \max (\|x_1\|, \|y_1\|) \\
\equiv \|(x, y)\| + \|(x_1, y_1)\|
\]
Let \( x \in U_y \). Then \( (x, y) \in U \) and so there exists \( r > 0 \) such that

\[
B((x, y), r) \in U.
\]

This says that if \( (u, v) \in X \times Y \) such that \( ||(u, v) - (x, y)|| < r \), then \( (u, v) \in U \). Thus if

\[
||(u, y) - (x, y)|| = ||u - x|| < r,
\]

then \( (u, y) \in U \). This has just said that \( B(x, r) \), the ball taken in \( X \) is contained in \( U_y \). This proves the lemma. ■

Or course one could also consider

\[
U_x \equiv \{ y : (x, y) \in U \}
\]

in the same way and conclude this set is open in \( Y \). Also, the generalization to many factors yields the same conclusion. In this case, for \( x \in \prod_{i=1}^n X_i \), let

\[
||x|| \equiv \max\left(||x_i||_{X_i} : x = (x_1, \ldots, x_n)\right)
\]

Then a similar argument to the above shows this is a norm on \( \prod_{i=1}^n X_i \). Consider the triangle inequality.

\[
|| (x_1, \ldots, x_n) + (y_1, \ldots, y_n) || \leq \max_i \left( ||x_i + y_i||_{X_i} \right) \leq \max_i \left( ||x_i||_{X_i} + ||y_i||_{X_i} \right)
\]

Corollary 17.8.2 Let \( U \subseteq \prod_{i=1}^n X_i \) be an open set and let

\[
U_{(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)} \equiv \{ x \in \mathbb{F}^n : (x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) \in U \}.
\]

Then \( U_{(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)} \) is an open set in \( \mathbb{F}^n \).

Proof: Let \( z \in U_{(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)} \). Then \( (x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n) \equiv x \in U \) by definition. Therefore, since \( U \) is open, there exists \( r > 0 \) such that \( B(x, r) \subseteq U \). It follows that for \( B(z, r)_{X_i} \) denoting the ball in \( X_i \), it follows that \( B(z, r)_{X_i} \subseteq U_{(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)} \) because to say that \( ||z - w||_{X_i} < r \) is to say that

\[
|| (x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n) - (x_1, \ldots, x_{i-1}, w, x_{i+1}, \ldots, x_n) || < r
\]

and so \( w \in U_{(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)} \). ■

Next is a generalization of the partial derivative.

Definition 17.8.3 Let \( g : U \subseteq \prod_{i=1}^n X_i \rightarrow Y \), where \( U \) is an open set. Then the map

\[
z \mapsto g(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n)
\]

is a function from the open set in \( X_i \),

\[
\{ z : x = (x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n) \in U \}
\]

to \( Y \). When this map is differentiable, its derivative is denoted by \( D_i g (x) \). To aid in the notation, for \( v \in X_i \), let \( \theta_i v \in \prod_{i=1}^n X_i \) be the vector \( (0, \ldots, v, \ldots, 0) \) where the \( v \) is in the \( i \)-th slot and for \( v \in \prod_{i=1}^n X_i \), let \( v_i \) denote the entry in the \( i \)-th slot of \( v \). Thus, by saying

\[
z \mapsto g(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n)
\]

is differentiable is meant that for \( v \in X_i \), sufficiently small,

\[
g(x + \theta_i v) - g(x) = D_i g (x) v + o(v).
\]

Note \( D_i g (x) \in \mathcal{L}(X_i, Y) \).
Definition 17.8.4 Let $U \subseteq X$ be an open set. Then $f : U \to Y$ is $\mathcal{C}^1 (U)$ if $f$ is differentiable and the mapping

$$x \to Df (x),$$

is continuous as a function from $U$ to $\mathcal{L}(X,Y)$.

With this definition of partial derivatives, here is the major theorem. Note the resemblance with the matrix of the derivative of a function having values in $\mathbb{R}^m$ in terms of the partial derivatives.

Theorem 17.8.5 Let $g, U, \prod_{i=1}^n X_i$, be given as in Definition 17.8.4. Then $g$ is $\mathcal{C}^1 (U)$ if and only if $D_i g$ exists and is continuous on $U$ for each $i$. In this case, $g$ is differentiable and

$$D g (x) (v) = \sum_k D_k g(x) v_k$$

(17.8)

where $v = (v_1, \cdots, v_n)$.

Proof: Suppose then that $D_i g$ exists and is continuous for each $i$. Note that

$$\sum_{j=1}^n \theta_j v_j = (v_1, \cdots, v_k, 0, \cdots, 0).$$

Thus $\sum_{j=1}^n \theta_j v_j = v$ and define $\sum_{j=1}^0 \theta_j v_j \equiv 0$. Therefore,

$$g (x + v) - g (x) = \sum_{k=1}^n \left[ g \left( x + \sum_{j=1}^k \theta_j v_j \right) - g \left( x + \sum_{j=1}^{k-1} \theta_j v_j \right) \right]$$

(17.9)

Consider the terms in this sum.

$$g \left( x + \sum_{j=1}^k \theta_j v_j \right) - g \left( x + \sum_{j=1}^{k-1} \theta_j v_j \right) = g (x + \theta_k v_k) - g (x) +$$

(17.10)

$$\left( g \left( x + \sum_{j=1}^k \theta_j v_j \right) - g (x + \theta_k v_k) \right) - \left( g \left( x + \sum_{j=1}^{k-1} \theta_j v_j \right) - g (x) \right)$$

(17.11)

and the expression in (17.11) is of the form $h (v_k) - h (0)$ where for small $w \in X_k$,

$$h (w) = g \left( x + \sum_{j=1}^{k-1} \theta_j v_j + \theta_k w \right) - g (x + \theta_k w).$$

Therefore,

$$D h (w) = D_k g \left( x + \sum_{j=1}^{k-1} \theta_j v_j + \theta_k w \right) - D_k g (x + \theta_k w)$$

and by continuity, $||D h (w)|| < \varepsilon$ provided $||v||$ is small enough. Therefore, by Theorem 17.8.4, the mean value inequality, whenever $||v||$ is small enough,

$$||h (v_k) - h (0)|| \leq \varepsilon ||v||$$

which shows that since $\varepsilon$ is arbitrary, the expression in (17.11) is $o (v)$. Now in (17.11)

$$g (x + \theta_k v_k) - g (x) = D_k g (x) v_k + o (v_k) = D_k g (x) v_k + o (v).$$
Therefore, referring to (17.8),

$$g(x + v) - g(x) = \sum_{k=1}^{n} D_k g(x) v_k + o(v)$$

which shows that $Dg(x)$ exists and equals the formula given in (17.8). Also $x \to Dg(x)$ is continuous since each of the $D_k g(x)$ are.

Next suppose $g$ is $C^1$. I need to verify that $D_k g(x)$ exists and is continuous. Let $v \in X_k$ sufficiently small. Then

$$g(x + \theta_k v) - g(x) = Dg(x) \theta_k v + o(\theta_k v) = Dg(x) \theta_k v + o(v)$$

since $||\theta_k v|| = ||v||$. Then $D_k g(x)$ exists and equals

$$Dg(x) \circ \theta_k$$

Now $x \to Dg(x)$ is continuous. Since $\theta_k$ is linear, it follows from Theorem 17.8.2 that $\theta_k : X_k \to \prod_{i=1}^{n} X_i$ is also continuous.

Note that the above argument also works at a single point $x$. That is, continuity at $x$ of the partials implies $Dg(x)$ exists and is continuous at $x$.

The way this is usually used is in the following corollary which has already been obtained. Remember the matrix of $Df(x)$. Recall that if a function is $C^1$ in the sense that $x \to Df(x)$ is continuous then all the partial derivatives exist and are continuous. The next corollary says that if the partial derivatives do exist and are continuous, then the function is differentiable and has continuous derivative.

**Corollary 17.8.6** Let $U$ be an open subset of $\mathbb{F}^n$ and let $f : U \to \mathbb{F}^m$ be $C^1$ in the sense that all the partial derivatives of $f$ exist and are continuous. Then $f$ is differentiable and

$$f(x + v) = f(x) + \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x) v_k + o(v).$$

Similarly, if the partial derivatives up to order $k$ exist and are continuous, then the function is $C^k$ in the sense that the first $k$ derivatives exist and are continuous.

### 17.9 Mixed Partial Derivatives

The following theorem about equality of partial derivatives was known to Euler around 1734 and was proved later.

**Theorem 17.9.1** Suppose $f : U \subseteq \mathbb{F}^2 \to \mathbb{R}$ where $U$ is an open set on which $f_x, f_y, f_{xy}$ and $f_{yx}$ exist. Then if $f_{xy}$ and $f_{yx}$ are continuous at the point $(x, y) \in U$, it follows

$$f_{xy}(x, y) = f_{yx}(x, y).$$

**Proof:** Since $U$ is open, there exists $r > 0$ such that $B((x, y), r) \subseteq U$. Now let $|t|, |s| < r/2$, $t, s$ real numbers and consider

$$\Delta(s, t) \equiv \frac{1}{st} \left\{ f(x + t, y + s) - f(x + t, y) - (f(x, y + s) - f(x, y)) \right\}. \quad (17.12)$$

Note that $(x + t, y + s) \in U$ because

$$|(x + t, y + s) - (x, y)| = |(t, s)| = \left( t^2 + s^2 \right)^{1/2} \leq \left( \frac{r^2}{4} + \frac{r^2}{4} \right)^{1/2} = \frac{r}{\sqrt{2}} < r.$$
As implied above, \( h(t) = f(x + t, y + s) - f(x + t, y) \). Therefore, by the mean value theorem from one variable calculus and the (one variable) chain rule,

\[
\Delta(s, t) = \frac{1}{s t} (h(t) - h(0)) = \frac{1}{s t} h'(\alpha t) t = \frac{1}{s} \left( f_x(x + \alpha t, y + s) - f_x(x + \alpha t, y) \right)
\]

for some \( \alpha \in (0, 1) \). Applying the mean value theorem again,

\[
\Delta(s, t) = f_{xy}(x + \alpha t, y + \beta s)
\]

where \( \alpha, \beta \in (0, 1) \).

If the terms \( f(x + t, y) \) and \( f(x, y + s) \) are interchanged in \( \Delta(s, t) \), \( \Delta(s, t) \) is unchanged and the above argument shows there exist \( \gamma, \delta \in (0, 1) \) such that

\[
\Delta(s, t) = f_{yx}(x + \gamma t, y + \delta s).
\]

Letting \((s, t) \to (0, 0)\) and using the continuity of \( f_{xy} \) and \( f_{yx} \) at \((x, y)\),

\[
\lim_{(s, t) \to (0,0)} \Delta(s, t) = f_{xy}(x, y) = f_{yx}(x, y).
\]

The following is obtained from the above by simply fixing all the variables except for the two of interest.

**Corollary 17.9.2** Suppose \( U \) is an open subset of \( X \) and \( f: U \to \mathbb{R} \) has the property that for two indices, \( k, l \), \( f_{kl}, f_{x_k}, f_{x_l}, f_{x_kx_l}, \) and \( f_{x_lx_k} \) exist on \( U \) and \( f_{x_kx_l} \) and \( f_{x_lx_k} \) are both continuous at \( x \in U \). Then \( f_{x_kx_l}(x) = f_{x_lx_k}(x) \).

By considering the real and imaginary parts of \( f \) in the case where \( f \) has values in \( \mathbb{C} \) you obtain the following corollary.

**Corollary 17.9.3** Suppose \( U \) is an open subset of \( \mathbb{F}^n \) and \( f: U \to \mathbb{F} \) has the property that for two indices, \( k, l \), \( f_{kl}, f_{x_k}, f_{x_l}, f_{x_kx_l}, \) and \( f_{x_lx_k} \) exist on \( U \) and \( f_{x_kx_l} \) and \( f_{x_lx_k} \) are both continuous at \( x \in U \). Then \( f_{x_kx_l}(x) = f_{x_lx_k}(x) \).

Finally, by considering the components of \( f \) you get the following generalization.

**Corollary 17.9.4** Suppose \( U \) is an open subset of \( \mathbb{F}^n \) and \( f: U \to \mathbb{F}^m \) has the property that for two indices, \( k, l \), \( f_{kl}, f_{x_k}, f_{x_l}, f_{x_kx_l}, \) and \( f_{x_lx_k} \) exist on \( U \) and \( f_{x_kx_l} \) and \( f_{x_lx_k} \) are both continuous at \( x \in U \). Then \( f_{x_kx_l}(x) = f_{x_lx_k}(x) \).

It is necessary to assume the mixed partial derivatives are continuous in order to assert they are equal. The following is a well known example.

**Example 17.9.5** Let

\[
f(x, y) = \begin{cases} 
\frac{xy(x^2 - y^2)}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0) \\
0 & \text{if } (x, y) = (0, 0)
\end{cases}
\]

From the definition of partial derivatives it follows immediately that \( f_x(0, 0) = f_y(0, 0) = 0 \). Using the standard rules of differentiation, for \((x, y) \neq (0, 0)\),

\[
f_x = \frac{x^4 - y^4 + 4x^2y^2}{(x^2 + y^2)^2}, \quad f_y = \frac{x^4 - y^4 - 4x^2y^2}{(x^2 + y^2)^2}
\]

Now

\[
f_{xy}(0, 0) = \lim_{y \to 0} \frac{f_x(0, y) - f_x(0, 0)}{y} = \lim_{y \to 0} \frac{-y^4}{(y^2)^2} = -1
\]
while
\[ f_{yx}(0, 0) = \lim_{x \to 0} \frac{f_y(x, 0) - f_y(0, 0)}{x} = \lim_{x \to 0} \frac{x^4}{(x^2)^2} = 1 \]

showing that although the mixed partial derivatives do exist at \((0, 0)\), they are not equal there.

Incidentally, the graph of this function appears very innocent. Its fundamental sickness is not apparent. It is like one of those whitened sepulchers mentioned in the Bible.

17.10 Newton’s Method

Remember Newton’s method from one variable calculus. It was an algorithm for finding the zeros of a function. Beginning with \(x_k\) the next iterate was

\[ x_{k+1} = x_k - \frac{f'(x_k)^{-1}}{f'(x_k)} \]

Of course the same thing can sometimes work in \(\mathbb{R}^n\) or even more generally. Here you have a function \(f(x)\) and you want to locate a zero. Then you could consider the sequence of iterates

\[ x_{k+1} = x_k - Df(x_k)^{-1} f(x_k) \]

If the sequence converges to \(x\) then you would have

\[ x = x - Df(x)^{-1} f(x) \]

and so you would need to have \(f(x) = 0\). In the next section, a modification of this well known method will be used to prove the Implicit function theorem. The modification is that you look for a solution to the equation near \(x_0\) and replace the above algorithm with the simpler one

\[ x_{k+1} = x_k - Df(x_0)^{-1} f(x_k) \]

Then if \(Tx = x - Df(x_0)^{-1} f(x)\), it follows that as long as \(x\) is sufficiently close to \(x_0\),

\[ DT(x) = I - Df(x_0)^{-1} Df(x) \]

and the norm of this transformation is very small so one can use the mean value inequality to conclude that \(T\) is a contraction mapping and provide a sequence of iterates which converge to a fixed point. Actually, the situation will be a little more complicated because we will do the implicit function theorem first, but this is the idea.
17.11 Exercises

1. For \((x, y) \neq (0, 0)\), let \(f(x, y) = \frac{xy^3}{x^2 + y^2}\). Show that this function has a limit as \((x, y) \to (0, 0)\) for \((x, y)\) on an arbitrary straight line through \((0, 0)\). Next show that this function fails to have a limit at \((0, 0)\).

2. Here are some scalar valued functions of several variables. Determine which of these functions are \(o(v)\). Here \(v\) is a vector in \(\mathbb{R}^n\), \(v = (v_1, \ldots, v_n)\).
   
   (a) \(v_1 v_2\)
   (b) \(v_2 \sin(v_1)\)
   (c) \(v_1^2 + v_2\)
   (d) \(v_2 \sin(v_1 + v_2)\)
   (e) \(v_1(v_1 + v_2 + xv_3)\)
   (f) \((e^{v_1} - 1 - v_1)\)
   (g) \((x \cdot v) |v|\)

3. Here is a function of two variables. \(f(x, y) = x^2y + x^2\). Find \(Df(x, y)\) directly from the definition. Recall this should be a linear transformation which results from multiplying by a \(1 \times 2\) matrix. Find this matrix.

4. Let \(f(x, y) = \begin{pmatrix} x^2 + y \\ y^2 \end{pmatrix}\). Compute the derivative directly from the definition. This should be the linear transformation which results from multiplying by a \(2 \times 2\) matrix. Find this matrix.

5. You have \(h(x) = g(f(x))\) Here \(x \in \mathbb{R}^n, f(x) \in \mathbb{R}^m\) and \(g(y) \in \mathbb{R}^p\) where \(f, g\) are appropriately differentiable. Thus \(Dh(x)\) results from multiplication by a matrix. Using the chain rule, give a formula for the \(ij^{th}\) entry of this matrix. How does this relate to multiplication of matrices? In other words, you have two matrices which correspond to \(Dg(f(x))\) and \(Df(x)\)
   
   Call \(z = g(y), y = f(x)\). Then

   \[Dg(y) = \begin{pmatrix} \frac{\partial z_1}{\partial y_1} & \cdots & \frac{\partial z_1}{\partial y_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_p}{\partial y_1} & \cdots & \frac{\partial z_p}{\partial y_m} \end{pmatrix}, \quad Df(x) = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_p}{\partial x_1} & \cdots & \frac{\partial y_p}{\partial x_n} \end{pmatrix}\]

   Explain the manner in which the \(ij^{th}\) entry of \(Dh(x)\) is

   \[\sum_k \frac{\partial z_i}{\partial y_k} \frac{\partial y_k}{\partial x_j}\]

   This is a review of the way we multiply matrices. what is the \(i^{th}\) row of \(Dg(y)\) and the \(j^{th}\) column of \(Df(x)\)?

6. Find \(f_x, f_y, f_z, f_{xy}, f_{yz}, f_{zy}\) for the following. Verify the mixed partial derivatives are equal.
   
   (a) \(x^2y^3z^4 + \sin(xyz)\)
   (b) \(\sin(xyz) + x^2yz\)

7. Suppose \(f\) is a continuous function and \(f : U \to \mathbb{R}\) where \(U\) is an open set and suppose that \(x \in U\) has the property that for all \(y\) near \(x\), \(f(x) \leq f(y)\). Prove that if \(f\) has all of its partial derivatives at \(x\), then \(f_x(x) = 0\) for each \(x_i\). \textbf{Hint:} Consider \(f(x + tv) = h(t)\). Argue that \(h'(0) = 0\) and then see what this implies about \(Df(x)\).
As an important application of Problem 7 consider the following. Experiments are done at \( n \) times, \( t_1, t_2, \cdots, t_n \) and at each time there results a collection of numerical outcomes. Denote by \( \{(t_i, x_i)\}_{i=1}^p \) the set of all such pairs and try to find numbers \( a \) and \( b \) such that the line \( x = at + b \) approximates these ordered pairs as well as possible in the sense that out of all choices of \( a \) and \( b \), \( \sum_{i=1}^p (at_i + b - x_i)^2 \) is as small as possible. In other words, you want to minimize the function of two variables \( f(a, b) \equiv \sum_{i=1}^p (at_i + b - x_i)^2 \). Find a formula for \( a \) and \( b \) in terms of the given ordered pairs. You will be finding the formula for the least squares regression line.

Let \( f \) be a function which has continuous derivatives. Show that \( u(t, x) = f(x - ct) \) solves the wave equation \( u_{tt} - c^2 u = 0 \). What about \( u(x, t) = f(x + ct) \)? Here \( \Delta u = u_{xx} \).

Show that if \( \Delta u = \lambda u \) where \( u \) is a function of only \( x \), then \( e^{\lambda t} u \) solves the heat equation \( u_t - \Delta u = 0 \). Here \( \Delta u = u_{xx} \).

Show that if \( f(x) = o(x) \), then \( f'(0) = 0 \).

Let \( f(x, y) \) be defined on \( \mathbb{R}^2 \) as follows. \( f(x, x^2) = 1 \) if \( x \neq 0 \). Define \( f(0, 0) = 0 \), and \( f(x, y) = 0 \) if \( y \neq x^2 \). Show that \( f \) is not continuous at \( (0, 0) \) but that
\[
\lim_{h \to 0} \frac{f(ha, hb) - f(0, 0)}{h} = 0
\]
for \((a, b)\) an arbitrary vector. Thus the Gateaux derivative exists at \((0, 0)\) in every direction but \( f \) is not even continuous there.

Let \( U \) be an open subset of \( \mathbb{R}^n \) and suppose that \( f : [a, b] \times U \to \mathbb{R} \) satisfies
\[
(x, y) \to \frac{\partial f}{\partial y_i}(x, y), (x, y) \to f(x, y)
\]
are all continuous. Show that
\[
\int_a^b f(x, y) \, dx, \int_a^b \frac{\partial f}{\partial y_i}(x, y) \, dx
\]
all make sense and that in fact
\[
\frac{\partial}{\partial y_i} \left( \int_a^b f(x, y) \, dx \right) = \int_a^b \frac{\partial f}{\partial y_i}(x, y) \, dx
\]
Also explain why
\[
y \to \int_a^b \frac{\partial f}{\partial y_i}(x, y) \, dx
\]
is continuous. **Hint:** You will need to use the theorems from one variable calculus about the existence of the integral for a continuous function. You may also want to use theorems about uniform continuity of continuous functions defined on compact sets.
15. I found this problem in Apostol’s book [III]. This is a very important result and is obtained very simply. Read it and fill in any missing details. Let

\[ g(x) = \int_0^1 \frac{e^{-x^2(1+t^2)}}{1+t^2} \, dt \]

and

\[ f(x) = \left( \int_0^x e^{-t^2} \, dt \right)^2. \]

Note

\[ \frac{\partial}{\partial x} \left( \frac{e^{-x^2(1+t^2)}}{1+t^2} \right) = -2xe^{-x^2(1+t^2)} \]

Explain why this is so. Also show the conditions of Problem [IV] are satisfied so that

\[ g'(x) = \int_0^1 \left( -2xe^{-x^2(1+t^2)} \right) \, dt. \]

Now use the chain rule and the fundamental theorem of calculus to find \( f'(x) \). Then change the variable in the formula for \( f'(x) \) to make it an integral from 0 to 1 and show

\[ f'(x) + g'(x) = 0. \]

Now this shows \( f(x) + g(x) \) is a constant. Show the constant is \( \pi/4 \) by letting \( x \to 0 \). Next take a limit as \( x \to \infty \) to obtain the following formula for the improper integral, \( \int_0^\infty e^{-t^2} \, dt \),

\[ \left( \int_0^\infty e^{-t^2} \, dt \right)^2 = \pi/4. \]

In passing to the limit in the integral for \( g \) as \( x \to \infty \) you need to justify why that integral converges to 0. To do this, argue the integrand converges uniformly to 0 for \( t \in [0, 1] \) and then explain why this gives convergence of the integral. Thus

\[ \int_0^\infty e^{-t^2} \, dt = \sqrt{\pi}/2. \]

16. The gamma function is defined for \( x > 0 \) as

\[ \Gamma(x) = \int_0^\infty e^{-t} t^{x-1} \, dt = \lim_{R \to \infty} \int_0^R e^{-t} t^{x-1} \, dt \]

Show this limit exists. Note you might have to give a meaning to

\[ \int_0^R e^{-t} t^{x-1} \, dt \]

if \( x < 1 \). Also show that

\[ \Gamma(x + 1) = x \Gamma(x), \quad \Gamma(1) = 1. \]

How does \( \Gamma(n) \) for \( n \) an integer compare with \( (n-1)! \)?

17. Show the mean value theorem for integrals. Suppose \( f \in C([a, b]) \). Then there exists \( x \in [a, b] \), in fact \( x \) can be taken in \((a, b)\), such that

\[ f(x)(b-a) = \int_a^b f(t) \, dt \]

You will need to recall simple theorems about the integral from one variable calculus.
18. In this problem is a short argument showing a version of what has become known as Fubini’s theorem. Suppose \( f \in C ([a, b] \times [c, d]) \). Then

\[
\int_a^b \int_c^d f(x, y) \, dy \, dx = \int_c^d \int_a^b f(x, y) \, dx \, dy
\]

First explain why the two iterated integrals make sense. **Hint:** To prove the two iterated integrals are equal, let \( a = x_0 < x_1 < \cdots < x_n = b \) and \( c = y_0 < y_1 < \cdots < y_m = d \) be two partitions of \([a, b]\) and \([c, d]\) respectively. Then explain why

\[
\int_a^b \int_c^d f(x, y) \, dy \, dx = \sum_{i=1}^{n} \sum_{j=1}^{m} \int_{x_{i-1}}^{x_i} \int_{y_{j-1}}^{y_j} f(s, t) \, dt \, ds
\]

\[
\int_c^d \int_a^b f(x, y) \, dx \, dy = \sum_{j=1}^{m} \sum_{i=1}^{n} \int_{y_{j-1}}^{y_j} \int_{x_{i-1}}^{x_i} f(s, t) \, ds \, dt
\]

Now use the mean value theorem for integrals to write

\[
\int_{x_{i-1}}^{x_i} \int_{y_{j-1}}^{y_j} f(s, t) \, dt \, ds = f(\hat{s}_i, \hat{t}_j) (x_i - x_{i-1}) (y_j - y_{j-1})
\]

do something similar for

\[
\int_{y_{j-1}}^{y_j} \int_{x_{i-1}}^{x_i} f(s, t) \, ds \, dt
\]

and then observe that the difference between the sums can be made as small as desired by simply taking suitable paritions.
Chapter 18

Implicit Function Theorem

18.1 Statement And Proof Of The Theorem

Recall the following notation. \( \mathcal{L}(X, Y) \) is the space of bounded linear mappings from \( X \) to \( Y \) where here \((X, \| \cdot \|_X)\) and \((Y, \| \cdot \|_Y)\) are normed linear spaces. Recall that this means that for each \( L \in \mathcal{L}(X, Y) \)

\[
\|L\| = \sup_{\|x\| \leq 1} \|Lx\| < \infty
\]

As shown earlier, this makes \( \mathcal{L}(X, Y) \) into a normed linear space. In case \( X \) is finite dimensional, \( \mathcal{L}(X, Y) \) is the same as the collection of linear maps from \( X \) to \( Y \). This was shown earlier. In what follows \( X, Y \) will be Banach spaces. If you like, think of them as finite dimensional normed linear spaces, but if you like more generality, just think: complete normed linear space and \( \mathcal{L}(X, Y) \) is the space of bounded linear maps.

Definition 18.1.1 A complete normed linear space is called a Banach space.

Theorem 18.1.2 If \( Y \) is a Banach space, then \( \mathcal{L}(X, Y) \) is also a Banach space.

Proof: Let \( \{L_n\} \) be a Cauchy sequence in \( \mathcal{L}(X, Y) \) and let \( x \in X \).

\[
\|L_nx - L_mx\| \leq \|x\| \|L_n - L_m\|
\]

Thus \( \{L_nx\} \) is a Cauchy sequence. Let

\[
Lx = \lim_{n \to \infty} L_nx.
\]

Then, clearly, \( L \) is linear because if \( x_1, x_2 \) are in \( X \), and \( a, b \) are scalars, then

\[
L(ax_1 + bx_2) = \lim_{n \to \infty} L_n(ax_1 + bx_2) = \lim_{n \to \infty} (aL_n x_1 + bL_n x_2) = aLx_1 + bLx_2.
\]

Also \( L \) is bounded. To see this, note that \( \{||L_n|||\} \) is a Cauchy sequence of real numbers because

\[
||L_n|| - ||L_m|| \leq ||L_n - L_m||.
\]

Hence there exists \( K > \sup\{||L_n|| : n \in \mathbb{N}\} \). Thus, if \( x \in X \),

\[
||Lx|| = \lim_{n \to \infty} ||L_n x|| \leq K||x||. \quad \blacksquare
\]

The following theorem is really nice. The series in this theorem is called the Neuman series.
Lemma 18.1.3 Let \((X, \|\cdot\|)\) is a Banach space, and if \(A \in \mathcal{L}(X, X)\) and \(\|A\| = r < 1\), then
\[
(I - A)^{-1} = \sum_{k=0}^{\infty} A^k \in \mathcal{L}(X, X)
\]
where the series converges in the Banach space \(\mathcal{L}(X, X)\). If \(O\) consists of the invertible maps in \(\mathcal{L}(X, X)\), then \(O\) is open and if \(\Im\) is the mapping which takes \(A\) to \(A^{-1}\), then \(\Im\) is continuous.

Proof: First of all, why does the series make sense?

\[
\left\| \sum_{k=p}^{q} A^k \right\| \leq \sum_{k=p}^{q} \| A^k \| \leq \sum_{k=p}^{q} \| A \|^k \leq \sum_{k=p}^{\infty} r^k \leq \frac{r^p}{1 - r}
\]

and so the partial sums are Cauchy in \(\mathcal{L}(X, X)\). Therefore, the series converges to something in \(\mathcal{L}(X, X)\) by completeness of this normed linear space. Now why is it the inverse?

\[
\sum_{k=0}^{\infty} A^k (I - A) = \lim_{n \to \infty} \sum_{k=0}^{n} A^k (I - A) = \lim_{n \to \infty} \left( \sum_{k=0}^{n} A^k - \sum_{k=1}^{n+1} A^k \right) = \lim_{n \to \infty} (I - A^{n+1}) = I
\]
because \(\|A^{n+1}\| \leq \|A\|^{n+1} \leq r^{n+1}\). Similarly,

\[
(I - A) \sum_{k=0}^{\infty} A^k = \lim_{n \to \infty} (I - A^{n+1}) = I
\]

and so this shows that this series is indeed the desired inverse.

Next suppose \(A \in O\) so \(A^{-1} \in \mathcal{L}(X, X)\). Then suppose \(\|A - B\| < \frac{r}{1 + \|A^{-1}\|}, r < 1\). Does it follow that \(B\) is also invertible?

\[
B = A - (A - B) = A \left[ I - A^{-1} (A - B) \right]
\]

Then \(\|A^{-1} (A - B)\| \leq \|A^{-1}\| \|A - B\| < r\) and so \([I - A^{-1} (A - B)]^{-1}\) exists. Hence

\[
B^{-1} = [I - A^{-1} (A - B)]^{-1} A^{-1}
\]

Thus \(O\) is open as claimed. As to continuity, let \(A, B\) be as just described. Then using the Neumann series,

\[
\|\Im A - \Im B\| = \left\| A^{-1} - [I - A^{-1} (A - B)]^{-1} A^{-1} \right\|
\]

\[
= \left\| A^{-1} - \sum_{k=0}^{\infty} \left( A^{-1} (A - B) \right)^k A^{-1} \right\| = \left\| \sum_{k=1}^{\infty} \left( A^{-1} (A - B) \right)^k A^{-1} \right\|
\]

\[
\leq \sum_{k=1}^{\infty} \| A^{-1} \|^{k+1} \| A - B \| = \| A - B \| \| A^{-1} \| \sum_{k=0}^{\infty} \| A^{-1} \|^k \left( \frac{r}{1 + \|A^{-1}\|} \right)^k
\]

\[
\leq \|B - A\| \|A^{-1}\|^2 \frac{1}{1 - r}.
\]

Thus \(\Im\) is continuous at \(A \in O\). ■

Lemma 18.1.4 Let
\[
O \equiv \{ A \in \mathcal{L}(X, Y) : A^{-1} \in \mathcal{L}(Y, X) \}
\]

and let
\[
\Im : O \to \mathcal{L}(Y, X), \ \Im A \equiv A^{-1}.
\]

Then \(O\) is open and \(\Im\) is in \(C^m(O)\) for all \(m = 1, 2, \cdots\). Also

\[
D\Im (A) (B) = -\Im (A) (B) \Im (A).
\]

(18.1)

In particular, \(\Im\) is continuous.
18.1. STATEMENT AND PROOF OF THE THEOREM

\[ \|B\| \leq \frac{1}{2} \|A^{-1}\|^{-1}. \]

Then
\[ \|A^{-1}B\| \leq \|A^{-1}\| \|B\| \leq \frac{1}{2} \]

and so by Lemma \[\text{Lemma}\],
\[ (I + A^{-1}B)^{-1} \in \mathcal{L}(X, X). \]

It follows that
\[ (A + B)^{-1} = (A + I)^{-1} = (I + A^{-1}B)^{-1} A^{-1} \in \mathcal{L}(Y, X). \]

Thus \( O \) is an open set.

Thus
\[ (A + B)^{-1} = \left( I + A^{-1}B \right)^{-1} A^{-1} = \sum_{n=0}^{\infty} (-1)^n (A^{-1}B)^n A^{-1} \]
\[ = [I - A^{-1}B + o(B)] A^{-1} \]

which shows that \( O \) is open and, also,
\[ \mathcal{J}(A + B) - \mathcal{J}(A) = \sum_{n=0}^{\infty} (-1)^n \left( (A^{-1}B)^n A^{-1} - A^{-1} \right) \]
\[ = -A^{-1}BA^{-1} + o(B) \]
\[ = -\mathcal{J}(A)(B) \mathcal{J}(A) + o(B) \]

which demonstrates \[\text{Lemma}\]. It follows from this that we can continue taking derivatives of \( \mathcal{J} \). For \( \|B_1\| \) small,
\[ -[D\mathcal{J}(A + B_1)(B) - D\mathcal{J}(A)(B)] = \]
\[ \mathcal{J}(A + B_1)(B) \mathcal{J}(A + B_1) - \mathcal{J}(A)(B) \mathcal{J}(A) \]
\[ = \mathcal{J}(A + B_1)(B) \mathcal{J}(A + B_1) - \mathcal{J}(A)(B) \mathcal{J}(A + B_1) + \]
\[ \mathcal{J}(A)(B) [\mathcal{J}(A)(B_1) \mathcal{J}(A) + o(B_1)] \]
\[ = \mathcal{J}(A)(B_1) \mathcal{J}(A) + o(B_1)] \mathcal{J}(A)[A^{-1} - A^{-1}B_1A^{-1} + o(B_1)] + \]
\[ \mathcal{J}(A)(B) [\mathcal{J}(A)(B_1) \mathcal{J}(A) + o(B_1)] \]
\[ = \mathcal{J}(A)(B_1) \mathcal{J}(A) + \mathcal{J}(A)(B) \mathcal{J}(A)(B_1) \mathcal{J}(A) + o(B_1) \]

and so
\[ D^2\mathcal{J}(A)(B_1)(B) = \mathcal{J}(A)(B_1) \mathcal{J}(A)(B) \mathcal{J}(A) + \mathcal{J}(A)(B) \mathcal{J}(A)(B_1) \mathcal{J}(A) \]

which shows \( \mathcal{J} \) is \( C^2(O) \). Clearly we can continue in this way which shows \( \mathcal{J} \) is in \( C^m(O) \) for all \( m = 1, 2, \cdots \).}

Here are the two fundamental results presented earlier which will make it easy to prove the implicit function theorem. First is the fundamental mean value inequality.
Theorem 18.1.5 Suppose $U$ is an open subset of $X$ and $f : U \to Y$ has the property that $Df(x)$ exists for all $x$ in $U$ and that, $x + t(y - x) \in U$ for all $t \in [0, 1]$. (The line segment joining the two points lies in $U$.) Suppose also that for all points on this line segment,

$$||Df(x + t(y - x))|| \leq M.$$  

Then

$$||f(y) - f(x)|| \leq M |y - x|.$$  

Next recall the following theorem about fixed points of a contraction map. It was Corollary 11.1.4.

Corollary 18.1.6 Let $B$ be a closed subset of the complete metric space $(X, d)$ and let $f : B \to X$ be a contraction map

$$d(f(x), f(\hat{x})) \leq rd(x, \hat{x}), \quad r < 1.$$  

Also suppose there exists $x_0 \in B$ such that the sequence of iterates $\{f^n(x_0)\}_{n=1}^{\infty}$ remains in $B$. Then $f$ has a unique fixed point in $B$ which is the limit of the sequence of iterates. This is a point $x \in B$ such that $f(x) = x$. In the case that $B = B(x_0, \delta)$, the sequence of iterates satisfies the inequality

$$d(f^n(x_0), x_0) \leq \frac{d(x_0, f(x_0))}{1 - r}$$  

and so it will remain in $B$ if

$$\frac{d(x_0, f(x_0))}{1 - r} < \delta.$$  

The implicit function theorem deals with the question of solving, $f(x, y) = 0$ for $x$ in terms of $y$ and how smooth the solution is. It is one of the most important theorems in mathematics. The proof I will give holds with no change in the context of infinite dimensional complete normed vector spaces when suitable modifications are made on what is meant by $L(X, Y)$. There are also even more general versions of this theorem than to normed vector spaces.

Recall that for $X,Y$ normed vector spaces, the norm on $X \times Y$ is of the form

$$||(x, y)|| = \max(||x||, ||y||).$$  

Theorem 18.1.7 (Implicit function theorem) Let $X, Y, Z$ be finite dimensional normed vector spaces and suppose $U$ is an open set in $X \times Y$. Let $f : U \to Z$ be in $C^1(U)$ and suppose

$$f(x_0, y_0) = 0, \quad D_1f(x_0, y_0)^{-1} \in L(Z, X).$$  

Then there exist positive constants, $\delta, \eta$, such that for every $y \in B(y_0, \eta)$ there exists a unique $x(y) \in B(x_0, \delta)$ such that

$$f(x(y), y) = 0.$$  

Furthermore, the mapping, $y \to x(y)$ is in $C^1(B(y_0, \eta))$.

Proof: Let $T(x, y) = x - D_1f(x_0, y_0)^{-1}f(x, y)$. Therefore,

$$D_1T(x, y) = I - D_1f(x_0, y_0)^{-1}D_1f(x, y).$$  

by continuity of the derivative which implies continuity of $D_1T$, it follows there exists $\delta > 0$ such that if $||x - x_0|| < \delta$ and $||y - y_0|| < \delta$, then

$$||D_1T(x, y)|| < \frac{1}{2}, \quad D_1f(x, y)^{-1} exists$$  

The second claim follows from Lemma 18.1.4. By the mean value inequality, Theorem 18.1.6, whenever $x, x' \in B(x_0, \delta)$ and $y \in B(y_0, \delta),

$$||T(x, y) - T(x', y)|| \leq \frac{1}{2} ||x - x'||.$$  


18.1. STATEMENT AND PROOF OF THE THEOREM

Also, it can be assumed \( \delta \) is small enough that for some \( M \) and all such \( (x,y) \),

\[
\left\| D_1 f (x_0, y_0)^{-1} \right\| \| D_2 f (x, y) \| < M \tag{18.7}
\]

Next, consider only \( y \) such that \( \| y - y_0 \| < \eta \) where \( \eta \) is so small that

\[
\| T (x_0, y) - x_0 \| < \frac{\delta}{3}
\]

Then for such \( y \), consider the mapping \( T_y (x) = T (x, y) \). Thus by Corollary 18.1.6. for each \( n \in \mathbb{N} \),

\[
\delta > \frac{2}{3} \delta \geq \frac{\| T_y (x_0) - x_0 \|}{1 - (1/2)} \geq \| T^n (x_0) - x_0 \|
\]

Then by 18.1.5. the sequence of iterations of this map \( T_y \) converges to a unique fixed point \( x (y) \) in the ball \( B (x_0, \delta) \). Thus, from the definition of \( T \), \( f (x (y), y) = 0 \). This is the implicitly defined function.

Next we show that this function is Lipschitz continuous. For \( y, \tilde{y} \) in \( B (y_0, \eta) \),

\[
\| T (x, y) - T (x, \tilde{y}) \| = \| D_1 f (x_0, y_0)^{-1} f (x, y) - D_1 f (x_0, y_0)^{-1} f (x, \tilde{y}) \| \leq M \| y - \tilde{y} \|
\]

thanks to the above estimate and the mean value inequality, Theorem 18.1.4.1. Note how convexity of \( B (y_0, \eta) \) which says that the line segment joining \( y, \tilde{y} \) is contained in \( B (y_0, \eta) \) is important to use this theorem. Then from this,

\[
\| x (y) - x (\tilde{y}) \| = \| T (x (y), y) - T (x (\tilde{y}), \tilde{y}) \| \leq \| T (x (y), y) - T (x (y), \tilde{y}) \| + \| T (x (y), \tilde{y}) - T (x (\tilde{y}), \tilde{y}) \|
\]

\[
\leq M \| y - \tilde{y} \| + \frac{1}{2} \| x (y) - x (\tilde{y}) \|
\]

Hence,

\[
\| x (y) - x (\tilde{y}) \| \leq 2M \| y - \tilde{y} \| \tag{18.8}
\]

Finally consider the claim that this implicitly defined function is \( C^1 \).

\[
0 = f (x (y + u), y + u) - f (x (y), y) = D_1 f (x (y), y) (x (y + u) - x (y)) + D_2 f (x (y), y) u + o (x (y + u) - x (y), u) \tag{18.9}
\]

Consider the last term. \( o (x (y + u) - x (y), u) / \| u \| \) equals

\[
\begin{cases}
\frac{o (x (y + u) - x (y), u)}{\|u\|} \max (\|x (y + u) - x (y), u\|, \|u\|) & \text{if } \|x (y + u) - x (y), u\| \neq 0 \\
0 & \text{if } \|x (y + u) - x (y), u\| = 0
\end{cases}
\]

Now the Lipschitz condition just established shows that

\[
\frac{\max (\|x (y + u) - x (y), u\|, \|u\|)}{\|u\|}
\]

is bounded for nonzero \( u \) sufficiently small that \( y, y + u \in B (y_0, \eta) \). Therefore,

\[
\lim_{u \to 0} \frac{o (x (y + u) - x (y), u)}{\|u\|} = 0
\]

Then 18.1.4. shows that

\[
0 = D_1 f (x (y), y) (x (y + u) - x (y)) + D_2 f (x (y), y) u + o (u)
\]
Therefore, solving for \( x(y + u) - x(y) \), it follows that
\[
x(y + u) - x(y) = -D_1 f(x(y), y)^{-1} D_2 f(x(y), y) u + D_1 f(x(y), y)^{-1} o(u)
\]
and now, the continuity of the partial derivatives \( D_1 f,D_2 f \), continuity of the map \( A \rightarrow A^{-1} \), along with the continuity of \( y \rightarrow x(y) \) shows that \( y \rightarrow x(y) \) is \( C^1 \) with derivative equal to
\[
-D_1 f(x(y), y)^{-1} D_2 f(x(y), y) \quad \blacksquare
\]

The next theorem is a very important special case of the implicit function theorem known as the inverse function theorem. Actually one can also obtain the implicit function theorem from the inverse function theorem. It is done this way in [3], [4] and in [4].

**Theorem 18.1.8 (inverse function theorem)** Let \( x_0 \in U \), an open set in \( X \), and let \( f : U \rightarrow Y \) where \( X, Y \) are finite dimensional normed vector spaces. Suppose
\[
f \text{ is } C^1(U), \quad \text{ and } \quad Df(x_0)^{-1} \in \mathcal{L}(Y, X).
\]
Then there exist open sets \( W \), and \( V \) such that
\[
x_0 \in W \subseteq U , \quad \text{ (18.11)}
\]
\[
f : W \rightarrow V \text{ is one to one and onto}, \quad \text{ (18.12)}
\]
\[
f^{-1} \text{ is } C^1, \quad \text{ (18.13)}
\]

**Proof:** Apply the implicit function theorem to the function
\[
F(x, y) \equiv f(x) - y
\]
where \( y_0 \equiv f(x_0) \). Thus the function \( y \rightarrow x(y) \) defined in that theorem is \( f^{-1} \). Now let
\[
W \equiv B(x_0, \delta) \cap f^{-1}(B(y_0, \eta))
\]
and
\[
V \equiv B(y_0, \eta).
\]
This proves the theorem. \( \blacksquare \)

### 18.2 More Derivatives

When you consider a \( C^k \) function \( f \) defined on an open set \( U \), you obtain the following
\[
Df(x) \in \mathcal{L}(X,Y), D^2f(x) \in \mathcal{L}(X, \mathcal{L}(X,Y)), D^3f(x) \in \mathcal{L}(X, \mathcal{L}(X, \mathcal{L}(X,Y)))
\]
and so forth. Thus they can each be considered as a linear transformation with values in some vector space. When you consider the vector spaces, you see that these can also be considered as multilinear functions on \( X \) with values in \( Y \). Now consider the product of two linear transformations \( A(y)B(y)w \), where everything is given to make sense and here \( w \) is an appropriate vector. Then if each of these linear transformations can be differentiated, you would do the following simple computation.

\[
(A(y + u)B(y + u) - A(y)B(y))(w)
\]
\[
= (A(y + u)B(y + u) - A(y)B(y + u) + A(y)B(y + u) - A(y)B(y))(w)
\]
\[
= (DA(y)u + o(u))B(y + u) + A(y)(DB(y)u + o(u))(w)
\]
18.3. THE CASE OF $\mathbb{R}^n$

In many applications of the implicit function theorem, 

$$f : U \subseteq \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$$

and $f(x_0, y_0) = 0$ while $f$ is $C^1$. How can you recognize the condition of the implicit function theorem which says $D_1 f(x_0, y_0)^{-1}$ exists? This is really not hard. You recall the matrix of the transformation $D_1 f(x_0, y_0)$ with respect to the usual basis vectors is

$$
\begin{pmatrix}
  f_{1,1}(x_0, y_0) & \cdots & f_{1,n}(x_0, y_0) \\
  \vdots & & \vdots \\
  f_{n,1}(x_0, y_0) & \cdots & f_{n,n}(x_0, y_0)
\end{pmatrix}
$$

and so $D_1 f(x_0, y_0)^{-1}$ exists exactly when the determinant of the above matrix is nonzero. This is the condition to check. In the general case, you just need to verify $D_1 f(x_0, y_0)$ is one to one and this can also be accomplished by looking at the matrix of the transformation with respect to some bases on $X$ and $Z$.

18.4 Exercises

1. Let $A \in \mathcal{L}(X,Y)$. Let $f(x) = Ax$. Verify from the definition that $Df(x) = A$. What if $f(x) = y + Ax$? Note the similarity with functions of a single variable.

2. You have a level surface given by

$$f(x,y,z) = 0, \quad f \text{ is } C^1(U), (x,y,z) \in U,$$
The question is whether this deserves to be called a surface. Using the implicit function theorem, show that if \( f (x_0, y_0, z_0) = 0 \) and if

\[
\frac{\partial f}{\partial z} (x_0, y_0, z_0) \neq 0
\]

then in some open subset of \( \mathbb{R}^3 \), the relation \( f (x, y, z) = 0 \) can be “solved” for \( z \) getting say \( z = z(x, y) \) such that \( f(x, y, z(x, y)) = 0 \). What happens if \( \frac{\partial f}{\partial z} (x_0, y_0, z_0) = 0 \) or \( \frac{\partial f}{\partial y} (x_0, y_0, z_0) \neq 0 \)?

Explain why \( z \) is a \( C^1 \) map for \((x, y)\) in some open set.

3. Let \( \mathbf{x}(t) = (x(t), y(t), z(t))^T \) be a vector valued function defined for \( t \in (a, b) \). Then \( D\mathbf{x}(t) \in \mathcal{L}(\mathbb{R}, \mathbb{R}^3) \). We usually denote this simply as \( \mathbf{x}'(t) \). Thus, considered as a matrix, it is the \( 3 \times 1 \) matrix

\[
(x'(t), y'(t), z'(t))^T
\]

the \( T \) indicating that you take the transpose. Don’t worry too much about this. You can also consider this as a vector. What is the geometric significance of this vector? The answer is that this vector is tangent to the curve traced out by \( \mathbf{x}(t) \) for \( t \in (a, b) \). Explain why this is so using the definition of the derivative. You need to describe what is meant by being tangent first. By saying that the line \( \mathbf{x} = \mathbf{a} + t\mathbf{b} \) is tangent to a parametric curve consisting of points traced out by \( \mathbf{x}(t) \) for \( t \in (-\delta, \delta) \) at the point \( \mathbf{a} = \mathbf{x}(t) \) which is on both the line and the curve, you would want to have

\[
\lim_{u \to 0} \frac{\mathbf{x}(t + u) - (\mathbf{a} + t\mathbf{b})}{u} = \mathbf{0}
\]

With this definition of what it means for a line to be tangent, explain why the line \( \mathbf{x}(t) + \mathbf{x}'(t) u \) for \( u \in (-\delta, \delta) \) is tangent to the curve determined by \( t \to \mathbf{x}(t) \) at the point \( \mathbf{x}(t) \). So why would you take the above as a definition of what it means to be tangent? Consider the component functions of \( \mathbf{x}(t) \). What does the above limit say about the component functions and the corresponding components of \( \mathbf{b} \) in terms of slopes of lines tangent to curves?

4. Let \( f(x, y, z) \) be a \( C^1 \) function \( f : U \to \mathbb{R} \) where \( U \) is an open set in \( \mathbb{R}^3 \). The gradient vector, defined as

\[
\left( \frac{\partial f}{\partial x}(x, y, z), \frac{\partial f}{\partial y}(x, y, z), \frac{\partial f}{\partial z}(x, y, z) \right)^T
\]

has fundamental geometric significance illustrated by the following picture.

The way we present this in engineering math is to consider a smooth \( C^1 \) curve \( (x(t), y(t), z(t)) \) for \( t \in (a, b) \) such that when \( t = c \in (a, b) \), \( (x(c), y(c), z(c)) \) equals the point \( (x, y, z) \) in the level surface and such that \( (x(t), y(t), z(t)) \) lies in this surface. Then

\[
0 = f(x(t), y(t), z(t))
\]

Show, using the chain rule, that the gradient vector at the point \( (x, y, z) \) is perpendicular to \( (x'(c), y'(c), z'(c)) \). Recall that the chain rule says that for

\[
h(t) = f(x(t), y(t), z(t)), Dh(t) =
\]
This problem illustrates what can happen when the gradient of a scalar valued function vanishes. Let

\[
\left( \frac{\partial f}{\partial x} (x(t), y(t), z(t)) \quad \frac{\partial f}{\partial y} (x(t), y(t), z(t)) \quad \frac{\partial f}{\partial z} (x(t), y(t), z(t)) \right) \begin{pmatrix} x'(t) \\ y'(t) \\ z'(t) \end{pmatrix}
\]

Since this holds for all such smooth curves in the surface which go through the given point, we say that the gradient vector is perpendicular to the surface. In the picture, there are two intersecting curves which are shown to intersect at a point of the surface. We present this to the students in engineering math and everyone is happy with it, but the argument is specious. Why do there exist any smooth curves in the surface through a point? What would you need to assume to justify the existence of smooth curves in the surface at some point of the level surface? Why?

5. This problem illustrates what can happen when the gradient of a scalar valued function vanishes or is not well defined. Consider the level surface given by \( z = \sqrt{x^2 + y^2} = 0 \). Sketch the graph of this surface. Why is there no unique tangent plane at the origin \((0, 0, 0)\)? Next consider \( z^2 = (x^2 + y^2) = 0 \). What about a well defined tangent plane at \((0, 0, 0)\)?

6. Suppose you have two level surfaces \( f(x, y, z) = 0 \) and \( g(x, y, z) = 0 \) which intersect at a point \((x_0, y_0, z_0)\), each \( f, g \) is \( C^1 \). Use the implicit function theorem to give conditions which will guarantee that the intersection of these two surfaces near this point is a curve. Explain why.

7. Let \( X, Y \) be Banach spaces and let \( U \) be an open subset of \( X \). Let \( f : U \to Y \) be \( C^1 (U) \), let \( x_0 \in U \), and \( \delta > 0 \) be given. Show there exists \( \varepsilon > 0 \) such that if \( x_1, x_2 \in B(x_0, \varepsilon) \), then

\[
\|f(x_1) - f(x_2) - Df(x_0)(x_1 - x_2)\| \leq \delta \|x_1 - x_2\|
\]

**Hint:** You know \( f(x_1) - f(x_2) = Df(x_2)(x_1 - x_2) + o(x_1 - x_2) \). Use continuity.

8. †This problem illustrates how if \( Df(x_0) \) is one to one, then near \( x_0 \) the same is true of \( f \). Suppose in this problem that all normed linear spaces are finite dimensional. Suppose \( Df(x_0) \) is one to one. Here \( f : U \to Y \) where \( U \subseteq X \).

(a) Show that there exists \( r > 0 \) such that \( \|Df(x_0)x\| \geq r \|x\| \). To do this, recall equivalence of norms.

(b) Use the above problem to show that there is \( \varepsilon > 0 \) such that \( f \) is one to one on \( B(x_0, \varepsilon) \) provided \( Df(x_0) \) is one to one.

9. If \( U, V \) are open sets in Banach spaces \( X, Y \) respectively and \( f : U \to V \) is one to one and onto and both \( f, f^{-1} \) are \( C^1 \), show that \( Df(x) : X \to Y \) is one to one and onto for each \( x \in U \).

**Hint:** \( f \circ f^{-1} = \text{identity} \). Now use chain rule.

10. A function \( f : U \subseteq \mathbb{C} \to \mathbb{C} \) where \( U \) is an open set subset of the complex numbers \( \mathbb{C} \) is called analytic if

\[
\lim_{h \to 0} \frac{f(z + h) - f(z)}{h} \equiv f'(z), \ z = x + iy
\]

exists and \( z \to f'(z) \) is continuous. Show that if \( f \) is analytic on an open set \( U \) and if \( f'(z) \neq 0 \), then there is an open set \( V \) containing \( z \) such that \( f(V) \) is open, \( f \) is one to one, and \( f, f^{-1} \) are both continuous. **Hint:** This follows very easily from the inverse function theorem. Recall that we have allowed for the field of scalars the complex numbers.

11. Problem 3 has to do with concluding that \( f \) is locally one to one if \( Df(x_0) \) is only known to be one to one. The next obvious question concerns the situation where \( Df(x_0) \) maybe is possibly not one to one but is onto. There are two parts, a linear algebra consideration, followed by an application of the inverse function theorem. Thus these two problems are each generalizations of the inverse function theorem.
(a) Suppose \( X \) is a finite dimensional vector space and \( M \in \mathcal{L}(X,Y) \) is onto \( Y \). Consider a basis for \( M(X) = Y, \{Mx_1, \ldots, Mx_n\} \). Verify that \( \{x_1, \ldots, x_n\} \) is linearly independent. Define \( X \equiv \text{span}(x_1, \ldots, x_n) \). Show that if \( M \) is the restriction of \( M \) to \( X \), then \( M \) is one to one and onto \( Y \).

(b) Now suppose \( f : U \subseteq X \rightarrow Y \) is \( C^1 \) and \( Df(x_0) \) is onto \( Y \). Show that there is a ball \( B(f(x_0), \varepsilon) \) and an open set \( V \subseteq X \) such that \( f(V) \supseteq B(f(x_0), \varepsilon) \) so that if \( Df(x) \) is onto for each \( x \in U \), then \( f(U) \) is an open set. This is called the open map theorem. You might use the inverse function theorem with the spaces \( X, Y \). This is a nice illustration of why we developed the inverse and implicit function theorems on arbitrary normed linear spaces. You will see that this is a fairly easy problem.

12. Recall that a function \( f : U \subseteq X \rightarrow Y \) where here assume \( X \) is finite dimensional, is Gateaux differentiable if
\[
\lim_{t \to 0} \frac{f(x + tv) - f(x)}{t} = D_v f(x)
\]
exists. Here \( t \in \mathbb{R} \). Suppose that \( x \to D_v f(x) \) exists and is continuous on \( U \). Show it follows that \( f \) is differentiable and in fact \( D_v f(x) = Df(x) v \). \textbf{Hint:} Let \( g(y) = f(\sum_i y_i x_i) \) and argue that the partial derivatives of \( g \) all exist and are continuous. Conclude that \( g \) is \( C^1 \) and then argue that \( f \) is just the composition of \( g \) with a linear map.

13. Let
\[
f(x, y) = \begin{cases} (x^2 - y^4)^2 & \text{if } (x, y) \neq (0, 0) \\ 1 & \text{if } (x, y) = (0, 0) \end{cases}
\]
Show that \( f \) is not differentiable, and in fact is not even continuous, but \( D_v f(0, 0) \) exists and equals 0 for every \( v \neq 0 \).

14. Let
\[
f(x, y) = \begin{cases} \frac{x y^4}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}
\]
Show that \( f \) is not differentiable, and in fact is not even continuous, but \( D_v f(0, 0) \) exists and equals 0 for every \( v \neq 0 \).

18.5 The Method Of Lagrange Multipliers

As an application of the implicit function theorem, consider the method of Lagrange multipliers from calculus. Recall the problem is to maximize or minimize a function subject to equality constraints. Let \( f : U \to \mathbb{R} \) be a \( C^1 \) function where \( U \subseteq \mathbb{R}^n \) and let
\[
g_i(x) = 0, \quad i = 1, \ldots, m
\]
be a collection of equality constraints with \( m < n \). Now consider the system of nonlinear equations
\[
f(x) = a
\]
\[
g_i(x) = 0, \quad i = 1, \ldots, m.
\]
x_0 is a local maximum if \( f(x_0) \geq f(x) \) for all \( x \) near \( x_0 \) which also satisfies the constraints. A local minimum is defined similarly. Let \( F : U \times \mathbb{R} \to \mathbb{R}^{m+1} \) be defined by
\[
F(x, a) = \begin{pmatrix} f(x) - a \\ g_1(x) \\ \vdots \\ g_m(x) \end{pmatrix}
\]
Now consider the \( m + 1 \times n \) Jacobian matrix, the matrix of the linear transformation, \( D_1 \mathbf{F} (x, a) \) with respect to the usual basis for \( \mathbb{R}^n \) and \( \mathbb{R}^{m+1} \):

\[
\begin{pmatrix}
  f_{x_1} (x_0) & \cdots & f_{x_n} (x_0) \\
  g_{1 x_1} (x_0) & \cdots & g_{1 x_n} (x_0) \\
  \vdots & & \vdots \\
  g_{m x_1} (x_0) & \cdots & g_{m x_n} (x_0)
\end{pmatrix}.
\]

If this matrix has rank \( m + 1 \) then some \( m + 1 \times m + 1 \) submatrix has nonzero determinant. It follows from the implicit function theorem that there exist \( m + 1 \) variables, \( x_{i_1}, \ldots, x_{i_{m+1}} \) such that the system

\[
\mathbf{F} (x, a) = 0 \quad (18.17)
\]

specifies these \( m + 1 \) variables as a function of the remaining \( n - (m + 1) \) variables and \( a \) in an open set of \( \mathbb{R}^{n-m} \). Thus there is a solution \( (x_0, a) \) to \( (18.18) \) for some \( x_0 \) close to \( x_0 \) whenever \( a \) is in some open interval. Therefore, \( x_0 \) cannot be either a local minimum or a local maximum. It follows that if \( x_0 \) is either a local maximum or a local minimum, then the above matrix must have rank less than \( m + 1 \) which requires the rows to be linearly dependent. Thus, there exist \( m \) scalars,

\[
\lambda_1, \ldots, \lambda_m,
\]

and a scalar \( \mu \), not all zero such that

\[
\mu \begin{pmatrix}
  f_{x_1} (x_0) \\
  \vdots \\
  f_{x_n} (x_0)
\end{pmatrix} = \lambda_1 \begin{pmatrix}
  g_{1 x_1} (x_0) \\
  \vdots \\
  g_{1 x_n} (x_0)
\end{pmatrix} + \cdots + \lambda_m \begin{pmatrix}
  g_{m x_1} (x_0) \\
  \vdots \\
  g_{m x_n} (x_0)
\end{pmatrix} \quad (18.18)
\]

If the column vectors

\[
\begin{pmatrix}
  g_{1 x_1} (x_0) \\
  \vdots \\
  g_{1 x_n} (x_0)
\end{pmatrix}, \ldots, \begin{pmatrix}
  g_{m x_1} (x_0) \\
  \vdots \\
  g_{m x_n} (x_0)
\end{pmatrix} \quad (18.19)
\]

are linearly independent, then, \( \mu \neq 0 \) and dividing by \( \mu \) yields an expression of the form

\[
\begin{pmatrix}
  f_{x_1} (x_0) \\
  \vdots \\
  f_{x_n} (x_0)
\end{pmatrix} = \lambda_1 \begin{pmatrix}
  g_{1 x_1} (x_0) \\
  \vdots \\
  g_{1 x_n} (x_0)
\end{pmatrix} + \cdots + \lambda_m \begin{pmatrix}
  g_{m x_1} (x_0) \\
  \vdots \\
  g_{m x_n} (x_0)
\end{pmatrix} \quad (18.20)
\]

at every point \( x_0 \) which is either a local maximum or a local minimum. This proves the following theorem.

**Theorem 18.5.1** Let \( U \) be an open subset of \( \mathbb{R}^n \) and let \( f : U \to \mathbb{R} \) be a \( C^1 \) function. Then if \( x_0 \in U \) is either a local maximum or local minimum of \( f \) subject to the constraints \( (18.18) \), then \( (18.18) \) must hold for some scalars \( \mu, \lambda_1, \cdots, \lambda_m \) not all equal to zero. If the vectors in \( (18.18) \) are linearly independent, it follows that an equation of the form \( (18.20) \) holds.

### 18.6 The Taylor Formula

First recall the Taylor formula with the Lagrange form of the remainder. It will only be needed on \( [0, 1] \) so that is what I will show.

**Theorem 18.6.1** Let \( h : [0, 1] \to \mathbb{R} \) have \( m + 1 \) derivatives. Then there exists \( t \in (0, 1) \) such that

\[
h (1) = h (0) + \sum_{k=1}^{m} \frac{h^{(k)} (0)}{k!} t^k + \frac{h^{(m+1)} (t)}{(m+1)!}.
\]
Proof: Let $K$ be a number chosen such that
\[ h(1) - \left( h(0) + \sum_{k=1}^{m} \frac{h^{(k)}(0)}{k!} + K \right) = 0 \]

Now the idea is to find $K$. To do this, let
\[ F(t) = h(1) - \left( h(t) + \sum_{k=1}^{m} \frac{h^{(k)}(t)}{k!} (1 - t)^k + K (1 - t)^{m+1} \right) \]
Then $F(1) = F(0) = 0$. Therefore, by Rolle’s theorem there exists $t$ between 0 and 1 such that $F'(t) = 0$. Thus,
\[ 0 = -F'(t) = h'(t) + \sum_{k=1}^{m} \frac{h^{(k+1)}(t)}{k!} (1 - t)^k - \sum_{k=1}^{m-1} \frac{h^{(k+1)}(t)}{k!} (1 - t)^k - K (m + 1) (1 - t)^m \]
And so
\[ h'(t) + \frac{h^{(m+1)}(t)}{m!} (1 - t)^m \]
and so
\[ K = \frac{h^{(m+1)}(t)}{(m + 1)!} \]
This proves the theorem. \[ \square \]

Now let $f : U \to \mathbb{R}$ where $U \subseteq X$ a normed vector space and suppose $f \in C^m(U)$ and suppose $D^{m+1} f(x)$ exists on $U$. Let $x \in U$ and let $r > 0$ be such that
\[ B(x, r) \subseteq U. \]
Then for $\|v\| < r$ consider
\[ f(x + tv) - f(x) = h(t) \]
for $t \in [0, 1]$. Then by the chain rule,
\[ h'(t) = Df(x+tv)(v), \quad h''(t) = D^2f(x+tv)(v)(v) \]
and continuing in this way,
\[ h^{(k)}(t) = D^{(k)}f(x+tv)(v)(v) \cdots (v) \equiv D^{(k)}f(x+tv)v^k. \]
It follows from Taylor’s formula for a function of one variable given above that
\[ f(x + v) = f(x) + \sum_{k=1}^{m} \frac{D^{(k)}f(x)v^k}{k!} + \frac{D^{(m+1)}f(x+tv)v^{m+1}}{(m + 1)!}. \tag{18.21} \]
This proves the following theorem.

**Theorem 18.6.2** Let $f : U \to \mathbb{R}$ and let $f \in C^{m+1}(U)$. Then if
\[ B(x, r) \subseteq U, \]
and $\|v\| < r$, there exists $t \in (0, 1)$ such that \[ \text{(18.21)} \] holds.
18.7 Second Derivative Test

Now consider the case where \( U \subseteq \mathbb{R}^n \) and \( f: U \to \mathbb{R} \) is \( C^2(U) \). Then from Taylor’s theorem, if \( v \) is small enough, there exists \( t \in (0, 1) \) such that

\[
f(x + v) = f(x) + Df(x)v + \frac{D^2f(x+tv)v^2}{2}.
\]

Consider

\[
D^2f(x+tv)(e_i)(e_j) \equiv D(D(f(x+tv))e_i)e_j = D\left(\frac{\partial f(x+tv)}{\partial x_i}\right)e_j = \frac{\partial^2 f(x+tv)}{\partial x_j \partial x_i}
\]

where \( e_i \) are the usual basis vectors. Letting

\[
v = \sum_{i=1}^{n} v_i e_i,
\]

the second derivative term in \( 18.22 \) reduces to

\[
\frac{1}{2} \sum_{i,j} D^2f(x+tv)(e_i)(e_j) v_i v_j = \frac{1}{2} \sum_{i,j} H_{ij}(x+tv) v_i v_j
\]

where

\[
H_{ij}(x+tv) = D^2f(x+tv)(e_i)(e_j) = \frac{\partial^2 f(x+tv)}{\partial x_j \partial x_i}.
\]

**Definition 18.7.1** The matrix whose \( ij \)th entry is \( \frac{\partial^2 f(x)}{\partial x_j \partial x_i} \) is called the Hessian matrix, denoted as \( H(x) \).

From Theorem 17.9.1, this is a symmetric real matrix, thus self adjoint. By the continuity of the second partial derivative,

\[
f(x + v) = f(x) + Df(x)v + \frac{1}{2} v^T H(x)v + \frac{1}{2} \left( v^T (H(x+tv) - H(x)) v \right).
\]

where the last two terms involve ordinary matrix multiplication and

\[
v^T = (v_1 \ldots v_n)
\]

for \( v_i \) the components of \( v \) relative to the standard basis.

**Definition 18.7.2** Let \( f: D \to \mathbb{R} \) where \( D \) is a subset of some normed vector space. Then \( f \) has a local minimum at \( x \in D \) if there exists \( \delta > 0 \) such that for all \( y \in B(x, \delta) \)

\[
f(y) \geq f(x).
\]

\( f \) has a local maximum at \( x \in D \) if there exists \( \delta > 0 \) such that for all \( y \in B(x, \delta) \)

\[
f(y) \leq f(x).
\]
Theorem 18.7.3 If \( f : U \to \mathbb{R} \) where \( U \) is an open subset of \( \mathbb{R}^n \) and \( f \) is \( C^2 \), suppose \( Df(x) = 0 \). Then if \( H(x) \) has all positive eigenvalues, \( x \) is a local minimum. If the Hessian matrix \( H(x) \) has all negative eigenvalues, then \( x \) is a local maximum. If \( H(x) \) has a positive eigenvalue, then there exists a direction in which \( f \) has a local minimum at \( x \), while if \( H(x) \) has a negative eigenvalue, there exists a direction in which \( H(x) \) has a local maximum at \( x \).

Proof: Since \( Df(x) = 0 \), formula 18.23 holds and by continuity of the second derivative, \( H(x) \) is a symmetric matrix. Thus \( H(x) \) has all real eigenvalues. Suppose first that \( H(x) \) has all positive eigenvalues and that all are larger than \( \delta^2 > 0 \). Then by Theorem 14.1.6, \( H(x) \) has an orthonormal basis of eigenvectors, \( \{v_i\}_{i=1}^n \) and if \( u \) is an arbitrary vector, such that \( u = \sum_{j=1}^n u_j v_j \) where \( u_j = u \cdot v_j \), then

\[
 u^T H(x) u = \sum_{j=1}^n u_j v_j^T H(x) \sum_{j=1}^n u_j v_j
 = \sum_{j=1}^n u_j^2 \lambda_j \geq \delta^2 \sum_{j=1}^n u_j^2 = \delta^2 |u|^2.
\]

From 18.23 and the continuity of \( H \), if \( v \) is small enough,

\[
f(x + v) \geq f(x) + \frac{1}{2} \delta^2 |v|^2 - \frac{1}{4} \delta^2 |v|^2 = f(x) + \frac{\delta^2}{4} |v|^2.
\]

This shows the first claim of the theorem. The second claim follows from similar reasoning. Suppose \( H(x) \) has a positive eigenvalue \( \lambda^2 \). Then let \( v \) be an eigenvector for this eigenvalue. Then from 18.24,

\[
f(x + tv) = f(x) + \frac{1}{2} t^2 v^T H(x) v + \frac{1}{2} t^2 (v^T (H(x + tv) - H(x)) v)
\]

which implies

\[
f(x + tv) \geq f(x) + \frac{1}{4} t^2 \lambda^2 |v|^2
\]

whenever \( t \) is small enough. Thus in the direction \( v \) the function has a local minimum at \( x \). The assertion about the local maximum in some direction follows similarly. This proves the theorem.

This theorem is an analogue of the second derivative test for higher dimensions. As in one dimension, when there is a zero eigenvalue, it may be impossible to determine from the Hessian matrix what the local qualitative behavior of the function is. For example, consider

\[
f_1(x, y) = x^4 + y^2, \quad f_2(x, y) = -x^4 + y^2.
\]

Then \( Df_1(0, 0) = 0 \) and for both functions, the Hessian matrix evaluated at \( (0,0) \) equals

\[
\begin{pmatrix}
0 & 0 \\
0 & 2
\end{pmatrix}
\]

but the behavior of the two functions is very different near the origin. The second has a saddle point while the first has a minimum there.
18.8 The Rank Theorem

This is a very interesting result. The proof follows Marsden and Hoffman. First here is some linear algebra.

**Theorem 18.8.1** Let \( L : \mathbb{R}^n \to \mathbb{R}^N \) have rank \( m \). Then there exists a basis \( \{u_1, \ldots, u_m, u_{m+1}, \ldots, u_n\} \) such that a basis for \( \ker (L) \) is \( \{u_{m+1}, \ldots, u_n\} \).

**Proof:** Since \( L \) has rank \( m \), there is a basis for \( L (\mathbb{R}^n) \) which is of the form

\[
\{Lu_1, \ldots, Lu_m\}
\]

Then if

\[
\sum_i c_i u_i = 0
\]

you can do \( L \) to both sides and conclude that each \( c_i = 0 \). Hence \( \{u_1, \ldots, u_m\} \) is linearly independent. Let \( \{v_1, \ldots, v_k\} \) be a basis for \( \ker (L) \). Let \( x \in \mathbb{R}^n \). Then \( Lx = \sum_{i=1}^m c_i Lu_i \) for some choice of scalars \( c_i \). Hence

\[
L \left( x - \sum_{i=1}^m c_i u_i \right) = 0
\]

which shows that there exist \( d_j \) such that

\[
x = \sum_{i=1}^m c_i u_i + \sum_{j=1}^k d_j v_j
\]

It follows that \( \text{span} (u_1, \ldots, u_m, v_1, \ldots, v_k) = \mathbb{R}^n \). Is this set of vectors linearly independent? Suppose

\[
\sum_{i=1}^m c_i u_i + \sum_{j=1}^k d_j v_j = 0
\]

Do \( L \) to both sides to get

\[
\sum_{i=1}^m c_i Lu_i = 0
\]

Thus each \( c_i = 0 \). Hence \( \sum_{j=1}^k d_j v_j = 0 \) and so each \( d_j = 0 \) also. It follows that \( k = n - m \) and we can let

\[
\{v_1, \ldots, v_k\} = \{u_{m+1}, \ldots, u_n\}. \blacksquare
\]

Another useful linear algebra result is the following lemma.

**Lemma 18.8.2** Let \( V \subseteq \mathbb{R}^n \) be a subspace and suppose \( A(x) \in L (V, \mathbb{R}^N) \) for \( x \) in some open set \( U \). Also suppose \( x \to A(x) \) is continuous for \( x \in U \). Then if \( A(x_0) \) is one to one on \( V \) for some \( x_0 \in U \), then it follows that for all \( x \) close enough to \( x_0 \), \( A(x) \) is also one to one on \( V \).

**Proof:** Consider \( V \) as an inner product space with the inner product from \( \mathbb{R}^n \) and \( A(x)^* A(x) \). Then \( A(x)^* A(x) \in L (V, V) \) and \( x \to A(x)^* A(x) \) is also continuous. Also for \( v \in V \),

\[
(A(x)^* A(x) v, v)_{\mathbb{R}^N} = (A(x) v, A(x) v)_{\mathbb{R}^N}
\]

If \( A(x_0)^* A(x_0) v = 0 \), then from the above, it follows that \( A(x_0) v = 0 \) also. Therefore, \( v = 0 \) and so \( A(x_0)^* A(x_0) \) is one to one on \( V \). For all \( x \) close enough to \( x_0 \), it follows from continuity that \( A(x)^* A(x) \) is also one to one. Thus, for such \( x \), if \( A(x) v = 0 \), then \( A(x)^* A(x) v = 0 \) and so \( v = 0 \). Thus, for \( x \) close enough to \( x_0 \), it follows that \( A(x) \) is also one to one on \( V \). \( \blacksquare \)
Theorem 18.8.3 Let \( f : A \subseteq \mathbb{R}^n \to \mathbb{R}^N \) where \( A \) is open in \( \mathbb{R}^n \). Let \( f \) be a \( C^r \) function and suppose that \( Df (x) \) has rank \( m \) for all \( x \in A \). Let \( x_0 \in A \). Then there are open sets \( U,V \subseteq \mathbb{R}^n \) with \( x_0 \in V \), and a \( C^r \) function \( h : U \to V \) with inverse \( h^{-1} : V \to U \) also \( C^r \) such that \( f \circ h \) depends only on \((x_1, \ldots, x_m)\).

Proof: Let \( L = Df (x_0) \), and \( N_0 = \ker L \). Using the above linear algebra theorem, there exists 
\( \{u_1, \ldots, u_m\} \)
such that \( \{Lu_1, \ldots, Lu_m\} \) is a basis for \( L \mathbb{R}^n \). Extend to form a basis for \( \mathbb{R}^n \), 
\( \{u_1, \ldots, u_m, u_{m+1}, \ldots, u_n\} \)
such that a basis for \( N_0 = \ker L \) is \( \{u_{m+1}, \ldots, u_n\} \). Let
\( M \equiv \text{span} (u_1, \ldots, u_m) \).

Let the coordinate maps be \( \psi_k \) so that if \( x \in \mathbb{R}^n \),
\( x = \psi_1 (x) u_1 + \cdots + \psi_n (x) u_n \)
Since these coordinate maps are linear, they are infinitely differentiable.

Next I will define coordinate maps for \( x \in \mathbb{R}^N \). Then by the above construction, \( \{Lu_1, \ldots, Lu_m\} \)
is a basis for \( L (\mathbb{R}^n) \). Let a basis for \( \mathbb{R}^N \) be
\( \{Lu_1, \ldots, Lu_m, v_{m+1}, \ldots, v_N\} \)
(Note that, since the rank of \( Df (x_0) \) is \( m \) you must have \( N \geq m \).) The coordinate maps \( \phi_i \) will be defined as follows for \( x \in \mathbb{R}^N \):
\( x = \phi_1 (x) Lu_1 + \cdots + \phi_m (x) Lu_m + \phi_{m+1} (x) v_{m+1} + \cdots + \phi_N (x) v_N \)
Now define two infinitely differentiable maps \( G : \mathbb{R}^n \to \mathbb{R}^n \) and \( H : \mathbb{R}^N \to \mathbb{R}^n \),
\( G (x) \equiv (0, \cdots, 0, \psi_{m+1} (x), \cdots, \psi_n (x)) \)
\( H (y) \equiv (\phi_1 (y), \cdots, \phi_m (y), 0, \cdots, 0) \)
For \( x \in A \subseteq \mathbb{R}^n \), let
\( g (x) \equiv H (f (x)) + G (x) \in \mathbb{R}^n \)
Thus the first term picks out the first \( m \) entries of \( f (x) \) and the second term the last \( n - m \) entries of \( x \). It is of the form
\( (\phi_1 (f (x)), \cdots, \phi_m (f (x)), \psi_{m+1} (x), \cdots, \psi_n (x)) \)
Then
\[ Dg (x_0) (v) = HL (v) + Gv = HLv + Gv \] (18.24)
which is of the form
\[ Dg (x_0) (v) = (\phi_1 (Lv), \cdots, \phi_m (Lv), \psi_{m+1} (v), \cdots, \psi_n (v)) \]
If this equals \( 0 \), then all the components of \( v \), \( \psi_{m+1} (v), \cdots, \psi_n (v) \) are equal to 0. Hence
\( v = \sum_{i=1}^{m} c_i u_i \).
But also the coordinates of \( Lv, \phi_1 (Lv), \cdots, \phi_m (Lv) \) are all zero so \( Lv = 0 \) and so \( 0 = \sum_{i=1}^{m} c_i Lu_i \) so by independence of the \( Lu_i \), each \( c_i = 0 \) and consequently \( v = 0 \).
This proves the conditions for the inverse function theorem are valid for \( g \). Therefore, there is an open ball \( U \) and an open set \( V \), \( x_0 \in V \), such that \( g : V \to U \) is a \( C^r \) map and its inverse \( g^{-1} : U \to V \) is also. We can assume by continuity and Lemma 18.8.2 that \( V \) and \( U \) are small enough that for each \( x \in V \), \( Dg(x) \) is one to one. This follows from the fact that \( x \to Dg(x) \) is continuous.

Since it is assumed that \( Df(x) \) is of rank \( m \), \( Df(x)(\mathbb{R}^n) \) is a subspace which is \( m \) dimensional, denoted as \( P_x \). Also denote \( L(\mathbb{R}^n) = L(M) \) as \( P \).

\[ \begin{array}{c}
\mathbb{R}^N \\
\uparrow \\
\downarrow \\
\downarrow \\
M \subseteq \mathbb{R}^n
\end{array} \]

Thus \( \{Lu_1, \ldots, Lu_m\} \) is a basis for \( P \). Using Lemma 18.8.2 again, by making \( V, U \) smaller if necessary, one can also assume that for each \( x \in V \), \( Df(x) \) is one to one on \( M \) (although not on \( \mathbb{R}^n \)) and \( HDf(x) \) is one to one on \( M \). This follows from continuity and the fact that \( L = Df(x_0) \) is one to one on \( M \). Therefore, it is also the case that \( Df(x) \) maps the \( m \) dimensional space \( M \) onto the \( m \) dimensional space \( P_x \) and \( H \) is one to one on \( P_x \). The reason for this last claim is as follows: If \( Hz = 0 \) where \( z \in P_x \), then \( HDf(x)w = 0 \) where \( w \in M \) and \( Df(x)w = z \). Hence \( w = 0 \) because \( HDf(x) \) is one to one, and so \( z = 0 \) which shows that indeed \( H \) is one to one on \( P_x \).

Denote as \( L_\infty \) the inverse of \( H \) which is defined on \( \mathbb{R}^m \times 0 \), \( L_\infty : \mathbb{R}^m \times 0 \to P_x \). That \( 0 \) refers to the \( N - m \) string of zeros in the definition given above for \( H \).

Define \( h \equiv g^{-1} \) and consider \( f_1 \equiv f \circ h \). It is desired to show that \( f_1 \) depends only on \( x_1, \ldots, x_m \). Let \( D_1 \) refer to \( (x_1, \ldots, x_m) \) and let \( D_2 \) refer to \( (x_{m+1}, \ldots, x_n) \). Then \( f = f_1 \circ g \) and so by the chain rule

\[ Df(x)(y) = Df_1(g(x))Dg(x)(y) \]  
(18.25)

Now as in 18.8.2, for \( y \in \mathbb{R}^n \),

\[ Dg(x)(y) = HDf(x)(y) + Gy \]

Recall that from the above definitions of \( H \) and \( G \),

\[ G(y) \equiv (0, \ldots, 0, \psi_{m+1}(y), \ldots, \psi_n(y)) \]

\[ H(Df(x)(y)) = (\phi_1(Df(x)y), \ldots, \phi_m(Df(x)y), 0, \ldots, 0) \]

Let \( \pi_1 : \mathbb{R}^n \to \mathbb{R}^m \) denote the projection onto the first \( m \) positions and \( \pi_2 \) the projection onto the last \( n - m \). Thus

\[ \pi_1Dg(x)(y) = (\phi_1(Df(x)y), \ldots, \phi_m(Df(x)y)) \]

\[ \pi_2Dg(x)(y) = (\psi_{m+1}(y), \ldots, \psi_n(y)) \]

Now in general, for \( z \in \mathbb{R}^n \),

\[ Df_1(g(x))z = D_1f_1(g(x))\pi_1z + D_2f_1(g(x))\pi_2z \]

Therefore, it follows that \( Df_1(g(x))Dg(x)(y) \) is given by

\[ Df(x)(y) = Df_1(g(x))Dg(x)(y) \]

\[ = D_1f_1(g(x))\pi_1Dg(x)(y) + D_2f_1(g(x))\pi_2Dg(x)(y) \]

\[ = \pi_1Dg(x)(y) \]

\[ Df(x)(y) = Df_1(g(x))Dg(x)(y) = D_1f_1(g(x))\pi_1HDf(x)(y) + D_2f_1(g(x))\pi_2Gy \]
We need to verify the last term equals 0. Solving for this term,

\[ D_2 f_1 (g(x)) \pi_2 G y = Df(x)(y) - D_1 f_1 (g(x)) \pi_1 H Df(x)(y) \]

As just explained, \( L_x \circ H \) is the identity on \( P_x \), the image of \( Df(x) \). Then

\[
D_2 f_1 (g(x)) \pi_2 G y = (L_x \circ H Df(x)(y) - D_1 f_1 (g(x)) \pi_1 H Df(x)(y)) (y)
\]

Factoring out that underlined term,

\[
D_2 f_1 (g(x)) \pi_2 G y = [L_x - D_1 f_1 (g(x)) \pi_1] H Df(x)(y)
\]

Now \( Df(x) : M \rightarrow P_x = Df(x)(\mathbb{R}^n) \) is onto. (This is based on the assumption that \( Df(x) \) has rank \( m \).) Thus it suffices to consider only \( y \in M \) in the right side of the above. However, for such \( y, \pi_2 G y = 0 \) because to be in \( M, \psi_k(y) = 0 \) if \( k \geq m + 1 \), and so the left side of the above equals 0. Thus it appears this term on the left is 0 for any \( y \) chosen. How can this be so? It can only take place if \( D_2 f_1 (g(x)) = 0 \) for every \( x \in V \). Thus, since \( g \) is onto, it can only take place if \( D_2 f_1 (x) = 0 \) for all \( x \in U \). Therefore on \( U \) it must be the case that \( f_1 \) depends only on \( x_1, \ldots, x_m \) as desired.

18.9 The Local Structure Of \( C^1 \) Mappings

In linear algebra it is shown that every invertible matrix can be written as a product of elementary matrices, those matrices which are obtained from doing a row operation to the identity matrix. Two of the row operations produce a matrix which will change exactly one entry of a vector when it is multiplied by the elementary matrix. The other row operation involves switching two rows and this has the effect of switching two entries in a vector when multiplied on the left by the elementary matrix. Thus, in terms of the effect on a vector, the mapping determined by the given matrix can be considered as a composition of mappings which either flip two entries of the vector or change exactly one. A similar local result is available for nonlinear mappings. I found this interesting result in the advanced calculus book by Rudin.

Definition 18.9.1 Let \( U \) be an open set in \( \mathbb{R}^n \) and let \( G : U \rightarrow \mathbb{R}^n \). Then \( G \) is called primitive if it is of the form

\[
G(x) = \left( x_1 \quad \cdots \quad \alpha(x) \quad \cdots \quad x_n \right)^T.
\]

Thus, \( G \) is primitive if it only changes one of the variables. A function \( F : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is called a flip if

\[
F(x_1, \ldots, x_k, \ldots, x_l, \ldots, x_n) = (x_1, \ldots, x_l, \ldots, x_k, \ldots, x_n)^T.
\]

Thus a function is a flip if it interchanges two coordinates. Also, for \( m = 1, 2, \ldots, n \), define

\[
P_m(x) = \left( x_1 \quad x_2 \quad \cdots \quad x_m \quad 0 \quad \cdots \quad 0 \right)^T.
\]

It turns out that if \( h(0) = 0, Dh(0)^{-1} \) exists, and \( h \) is \( C^1 \) on \( U \), then \( h \) can be written as a composition of primitive functions and flips. This is a very interesting application of the inverse function theorem.

Theorem 18.9.2 Let \( h : U \rightarrow \mathbb{R}^n \) be a \( C^1 \) function with \( h(0) = 0, Dh(0)^{-1} \) exists. Then there is an open set \( V \subseteq U \) containing 0, flips \( F_1, \ldots, F_{n-1} \), and primitive functions \( G_n, G_{n-1}, \ldots, G_1 \) such that for \( x \in V \),

\[
h(x) = F_1 \circ \cdots \circ F_{n-1} \circ G_n \circ G_{n-1} \circ \cdots \circ G_1(x).
\]

The primitive function \( G_j \) leaves \( x_i \) unchanged for \( i \neq j \).
Proof: Let

\[ h_1(x) \equiv h(x) = \begin{pmatrix} \alpha_1(x) & \cdots & \alpha_n(x) \end{pmatrix}^T \]

and

\[ Dh(0)e_1 = \begin{pmatrix} \alpha_{1,1}(0) & \cdots & \alpha_{n,1}(0) \end{pmatrix}^T \]

where \( \alpha_{k,1} \) denotes \( \frac{\partial \alpha_k}{\partial x_1} \). Since \( Dh(0) \) is one to one, the right side of this expression cannot be zero. Hence there exists some \( k \) such that \( \alpha_{k,1}(0) \neq 0 \). Now define

\[ G_1(x) \equiv \begin{pmatrix} \alpha_k(x) & x_2 & \cdots & x_n \end{pmatrix}^T \]

Then the matrix of \( DG_1(0) \) is of the form

\[ \begin{pmatrix} \alpha_{k,1}(0) & \cdots & \alpha_{k,n}(0) \\ 0 & 1 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \]

and its determinant equals \( \alpha_{k,1}(0) \neq 0 \). Therefore, by the inverse function theorem, there exists an open set \( U_1 \), containing \( 0 \) and an open set \( V_2 \) containing \( 0 \) such that \( G_1(U_1) = V_2 \) and \( G_1 \) is one to one and onto, such that it and its inverse are both \( C^1 \). Let \( F_1 \) denote the flip which interchanges \( x_k \) with \( x_1 \). Now define

\[ h_2(y) = F_1 \circ h_1 \circ G_1^{-1}(y) \]

Thus

\[ h_2(G_1(x)) = F_1 \circ h_1(x) = \begin{pmatrix} \alpha_k(x) & \cdots & \alpha_1(x) & \cdots & \alpha_n(x) \end{pmatrix}^T \quad (18.26) \]

Therefore,

\[ P_1 h_2(G_1(x)) = \begin{pmatrix} \alpha_k(x) & 0 & \cdots & 0 \end{pmatrix}^T. \]

Also

\[ P_1(G_1(x)) = \begin{pmatrix} \alpha_k(x) & 0 & \cdots & 0 \end{pmatrix}^T \]

so \( P_1 h_2(y) = P_1(y) \) for all \( y \in V_2 \). Also, \( h_2(0) = 0 \) and \( Dh_2(0)^{-1} \) exists because of the definition of \( h_2 \) above and the chain rule. Since \( F_1^2 = I \), the identity map, it follows from \( (18.26) \) that

\[ h(x) = h_1(x) = F_1 \circ h_2 \circ G_1(x). \quad (18.27) \]

Note that on an open set \( V_2 \equiv G_1(U_1) \) containing the origin, \( h_2 \) leaves the first entry unchanged. This is what \( P_1 h_2(G_1(x)) = P_1(G_1(x)) \) says. In contrast, \( h_1 = h \) left possibly no entries unchanged.

Suppose then, that for \( m \geq 2 \), \( h_m \) leaves the first \( m - 1 \) entries unchanged,

\[ P_{m-1} h_m(x) = P_{m-1}(x) \quad (18.28) \]

for all \( x \in U_m \), an open subset of \( U \) containing \( 0 \), and \( h_m(0) = 0 \), \( Dh_m(0)^{-1} \) exists. From \( (18.26) \), \( h_m(x) \) must be of the form

\[ h_m(x) = \begin{pmatrix} x_1 & \cdots & x_{m-1} & \alpha_1(x) & \cdots & \alpha_n(x) \end{pmatrix}^T \]

where these \( \alpha_k \) are different than the ones used earlier. Then

\[ Dh_m(0)e_m = \begin{pmatrix} 0 & \cdots & 0 & \alpha_{1,m}(0) & \cdots & \alpha_{n,m}(0) \end{pmatrix}^T \neq 0 \]
because \( D\mathbf{h}_m(0)^{-1} \) exists. Therefore, there exists a \( k \geq m \) such that \( \alpha_{k,m}(0) \neq 0 \), not the same \( k \) as before. Define

\[
\mathbf{G}_m(x) \equiv \begin{pmatrix} x_1 & \cdots & x_{m-1} & \alpha_k(x) & \cdots & x_n \end{pmatrix}^T
\]

so a change in \( \mathbf{G}_m \) occurs only in the \( m \)th slot. Then \( \mathbf{G}_m(0) = 0 \) and \( D\mathbf{G}_m(0)^{-1} \) exists similar to the above. In fact

\[
\det (D\mathbf{G}_m(0)) = \alpha_{k,m}(0).
\]

Therefore, by the inverse function theorem, there exists an open set \( V_{m+1} \) containing \( 0 \) such that \( V_{m+1} = \mathbf{G}_m(U_m) \) with \( \mathbf{G}_m \) and its inverse being one to one, continuous and onto. Let \( \mathbf{F}_m \) be the flip which flips \( x_m \) and \( x_k \). Then define \( \mathbf{h}_{m+1} \) on \( V_{m+1} \) by

\[
\mathbf{h}_{m+1}(y) = \mathbf{F}_m \circ \mathbf{h}_m \circ \mathbf{G}_m^{-1}(y).
\]

Thus for \( x \in U_m \),

\[
\mathbf{h}_{m+1}(\mathbf{G}_m(x)) = (\mathbf{F}_m \circ \mathbf{h}_m)(x).
\]

and consequently, since \( \mathbf{F}_m^2 = I \),

\[
\mathbf{F}_m \circ \mathbf{h}_{m+1} \circ \mathbf{G}_m(x) = \mathbf{h}_m(x)
\]

It follows

\[
P_m \mathbf{h}_{m+1}(\mathbf{G}_m(x)) = P_m(\mathbf{F}_m \circ \mathbf{h}_m)(x)
\]

\[
= \left( \begin{array}{cccc} x_1 & \cdots & x_{m-1} & \alpha_k(x) & 0 & \cdots & 0 \end{array} \right)^T
\]

and

\[
P_m(\mathbf{G}_m(x)) = \left( \begin{array}{cccc} x_1 & \cdots & x_{m-1} & \alpha_k(x) & 0 & \cdots & 0 \end{array} \right)^T.
\]

Therefore, for \( y \in V_{m+1} \),

\[
P_m \mathbf{h}_{m+1}(y) = P_m(y).
\]

As before, \( \mathbf{h}_{m+1}(0) = 0 \) and \( D\mathbf{h}_{m+1}(0)^{-1} \) exists. Therefore, we can apply (18.31) repeatedly, obtaining the following:

\[
\mathbf{h}(x) = \mathbf{F}_1 \circ \mathbf{h}_2 \circ \mathbf{G}_1(x)
\]

\[
= \mathbf{F}_1 \circ \mathbf{F}_2 \circ \mathbf{h}_3 \circ \mathbf{G}_2 \circ \mathbf{G}_1(x)
\]

\[
\vdots
\]

\[
= \mathbf{F}_1 \circ \cdots \circ \mathbf{F}_{n-1} \circ \mathbf{h}_n \circ \mathbf{G}_{n-1} \circ \cdots \circ \mathbf{G}_1(x)
\]

where \( \mathbf{h}_n \) fixes the first \( n-1 \) entries,

\[
P_{n-1} \mathbf{h}_n(x) = P_{n-1}(x) = \left( \begin{array}{cccc} x_1 & \cdots & x_{n-1} & 0 \end{array} \right)^T
\]

and so \( \mathbf{h}_n(x) \) is a primitive mapping of the form

\[
\mathbf{h}_n(x) = \left( \begin{array}{cccc} x_1 & \cdots & x_{n-1} & \alpha(x) \end{array} \right)^T.
\]

Therefore, define the primitive function \( \mathbf{G}_n(x) \) to equal \( \mathbf{h}_n(x) \).

18.10 Brouwer Fixed Point Theorem \( \mathbb{R}^n \)

This is on the Brouwer fixed point theorem and a discussion of some of the manipulations which are important regarding simplices. This here is an approach based on combinatorics or graph theory. It features the famous Sperner’s lemma. It uses very elementary concepts from linear algebra in an essential way.
18.10.1 Simplices And Triangulations

**Definition 18.10.1** Define an $n$ simplex, denoted by $[x_0, \cdots, x_n]$, to be the convex hull of the $n + 1$ points, $\{x_0, \cdots, x_n\}$ where $\{x_i - x_0\}_{i=1}^n$ are linearly independent. Thus

$$[x_0, \cdots, x_n] = \left\{ \sum_{i=0}^n t_i x_i : \sum_{i=0}^n t_i = 1, t_i \geq 0 \right\}.$$ 

Note that $\{x_j - x_m\}_{j \neq m}$ are also independent.

Since $\{x_i - x_0\}_{i=1}^n$ is independent, the $t_i$ are uniquely determined. If two of them are

$$\sum_{i=0}^n t_i x_i = \sum_{i=0}^n s_i x_i$$

Then

$$\sum_{i=0}^n t_i (x_i - x_0) = \sum_{i=0}^n s_i (x_i - x_0)$$

so $t_i = s_i$ for $i \geq 1$ by independence. Since the $s_i$ and $t_i$ sum to 1, it follows that also $s_0 = t_0$.

If $n \leq 2$, the simplex is a triangle, line segment, or point. If $n \leq 3$, it is a tetrahedron, triangle, line segment or point. To say that $\{x_i - x_0\}_{i=1}^n$ are independent is to say that $\{x_i - x_r\}_{i \neq r}$ are independent for each fixed $r$. Indeed, if $x_i - x_r = \sum_{j \neq i, r} c_j (x_j - x_r)$, then you would have

$$x_i - x_0 + x_0 - x_r = \sum_{j \neq i, r} c_j (x_j - x_0) + \left( \sum_{j \neq i, r} c_j \right) x_0$$

and it follows that $x_i - x_0$ is a linear combination of the $x_j - x_0$ for $j \neq i$, contrary to assumption.

A simplex $S$ can be triangulated. This means it is the union of smaller sub-simplices such that if $S_1, S_2$ are two simplices in the triangulation, with

$$S_1 \equiv [z_0^1, \cdots, z_m^1], \quad S_2 \equiv [z_0^2, \cdots, z_p^2]$$

then

$$S_1 \cap S_2 = [x_{k_0}, \cdots, x_{k_r}]$$

where $[x_{k_0}, \cdots, x_{k_r}]$ is in the triangulation and

$$\{x_{k_0}, \cdots, x_{k_r}\} = \{z_0^1, \cdots, z_m^1\} \cap \{z_0^2, \cdots, z_p^2\}$$

or else the two simplices do not intersect.

Does there exist a triangulation in which all sub-simplices have diameter less than $\varepsilon$? This is obvious if $n \leq 2$. Supposing it to be true for $n - 1$, is it also so for $n$? The barycenter $b$ of a simplex $[x_0, \cdots, x_n]$ is just $\frac{1}{n+1} \sum x_i$. This point is not in the convex hull of any of the faces, those simplices of the form $[x_0, \cdots, x_k, \cdots, x_n]$ where the hat indicates $x_k$ has been left out. Thus $[x_0, \cdots, b, \cdots, x_n]$ is a $n$ simplex also. Now in general, if you have an $n$ simplex $[x_0, \cdots, x_m]$, its diameter is the maximum of $|x_k - x_l|$ for all $k \neq l$. Consider $[b - x_j]$. It equals

$$\left| \sum_{i=0}^n \frac{1}{n+1} (x_i - x_0) \right| = \left| \sum_{i \neq j} \frac{1}{n+1} (x_i - x_j) \right| \leq \frac{n}{n+1} \text{diam} (S).$$

Next consider the $k$th face of $S [x_0, \cdots, x_k, \cdots, x_n]$. By induction, it has a triangulation into simplices which each have diameter no more than $\frac{n}{n+1} \text{diam} (S)$. Let these $n - 1$ simplices be denoted by $\{S_{i,k}^0, \cdots, S_{i,k}^m\}$. Then the simplices $\{S_{i,k}^n, b\}_{i=1,k=1}^{m,n+1}$ are a triangulation of $S$ such that $\text{diam} \left( S_{i,k}^n, b \right) \leq \frac{n}{n+1} \text{diam} (S)$. Do for $[S_{i,k}^n, b]$ what was just done for $S$ obtaining a triangulation of $S$ as the union of what is obtained such that each simplex has diameter no more than $\left(\frac{n}{n+1}\right)^2 \text{diam} (S)$. Continuing this way shows the existence of the desired triangulation.
18.10.2 Labeling Vertices

Next is a way to label the vertices. Let \( p_0, \ldots, p_n \) be the first \( n + 1 \) prime numbers. All vertices of a simplex \( S = [x_0, \ldots, x_n] \) having \( \{x_k - x_0\}_{k=1}^n \) independent will be labeled with one of these primes. In particular, the vertex \( x_k \) will be labeled as \( p_k \) if the simplex is \( [x_0, \ldots, x_n] \). The value of a simplex will be the product of its labels. Triangulate this \( S \). Consider a 1 simplex coming from the original simplex \( [x_{k_1}, x_{k_2}] \), label one end as \( p_{k_1} \), and the other as \( p_{k_2} \). Then label all other vertices of this triangulation which occur on \( [x_{k_1}, x_{k_2}] \) either \( p_{k_1} \) or \( p_{k_2} \). Note that by independence of \( \{x_k - x_0\}_{k \neq \ell} \), this cannot introduce an inconsistency because the segment cannot contain any other vertex of \( S \). Then obviously there will be an odd number of simplices in this triangulation having value \( p_{k_1} p_{k_2} \), that is a \( p_k \) at one end and a \( p_{k_2} \) at the other. Suppose that the labeling has been done for all vertices of the triangulation which are on \( [x_{j_1}, \ldots, x_{j_{k+1}}] \):

\[
\{x_{j_1}, \ldots, x_{j_{k+1}}\} \subseteq \{x_0, \ldots, x_n\}
\]

any \( k \) simplex for \( k \leq n - 1 \), and there is an odd number of simplices from the triangulation having value equal to \( \prod_{i=k+1}^{n} p_i \) on this simplex. Consider \( \hat{S} = [x_{j_1}, \ldots, x_{j_{k+1}}, x_{j_{k+2}}] \). Then by induction, there is an odd number of \( k \) simplices on the \( s^\text{th} \) face \( [x_{j_1}, \ldots, x_{j_s}, \ldots, x_{j_{k+1}}] \) having value \( \prod_{i \neq s} p_i \). In particular the face \( [x_{j_1}, \ldots, x_{j_{k+1}}, x_{j_{k+2}}] \) has an odd number of simplices with value \( \prod_{i=1}^{k+2} p_i \).

Now no simplex in any other face of \( \hat{S} \) can have this value by uniqueness of prime factorization. Label the “interior” vertices, those \( u \) having all \( s_i > 0 \) in \( u = \sum_{i=1}^{k+2} s_i x_{j_i} \), (These have not yet been labeled) with any of the \( p_{i_1}, \ldots, p_{i_{k+2}} \). Pick a simplex on the face \( [x_{j_1}, \ldots, x_{j_{k+1}}, x_{j_{k+2}}] \) which has value \( \prod_{i=k+1}^{n} p_i \) and cross this simplex into \( \hat{S} \). Continue crossing simplices having value \( \prod_{i=k+1}^{n} p_i \) which have not been crossed till the process ends. It must end because there are an odd number of these simplices having value \( \prod_{i=k+1}^{n} p_i \). If the process leads to the outside of \( \hat{S} \), then one can always enter it again because there are an odd number of simplices with value \( \prod_{i=1}^{k+1} p_i \) available and you will have used up an even number. When the process ends, the value of the simplex must be \( \prod_{i=1}^{k+2} p_i \) because it will have the additional label \( p_{k+2} \) on a vertex since if not, there will be another way out of the simplex. This identifies a simplex in the triangulation with value \( \prod_{i=1}^{k+2} p_i \).

Then repeat the process with \( \prod_{i=k+1}^{n} p_i \) valued simplices on \( [x_{j_1}, \ldots, x_{j_{k+1}}, x_{j_{k+2}}] \) which have not been crossed. Repeating the process, entering from the outside, cannot deliver a \( \prod_{i=1}^{k+2} p_i \) valued simplex encountered earlier. This is because you cross faces labeled \( \prod_{i=k+1}^{n} p_i \). If the remaining vertex is labeled \( p_{j_i} \) where \( i \neq k+2 \), then this yields exactly one other face to cross. There are two, the one with the first vertex \( p_{j_i} \) and the next one with the new vertex labeled \( p_{j_i} \) substituted for the first vertex having this label. Thus there is either one route in to a simplex or two. Thus, starting at a simplex labeled \( \prod_{i=k+1}^{n} p_i \) one can cross faces having this value till one is led to the \( \prod_{i=1}^{k+1} p_i \) valued simplex on the selected face of \( \hat{S} \). In other words, the process is one to one in selecting a \( \prod_{i=k+1}^{n} p_i \) vertex from crossing such a vertex on the selected face of \( \hat{S} \). Continue doing this, crossing a \( \prod_{i=k+1}^{n} p_i \) simplex on the face of \( \hat{S} \) which has not been crossed previously. This identifies an odd number of simplices having value \( \prod_{i=1}^{k+2} p_i \). These are the ones which are “accessible” from the outside using this process. If there are any which are not accessible from outside, applying the same process starting inside one of these, leads to exactly one other inaccessible simplex with value \( \prod_{i=1}^{k+2} p_i \). Hence these inaccessible simplices occur in pairs and so there are an odd number of simplices in the triangulation having value \( \prod_{i=1}^{k+2} p_i \). We refer to this procedure of labeling as Sperner’s lemma. The system of labeling is well defined thanks to the assumption that \( \{x_k - x_0\}_{k=1}^n \) is independent which implies that \( \{x_k - x_0\}_{k \neq \ell} \) is also linearly independent. Thus there can be no ambiguity in the labeling of vertices on any “face” the convex hull of some of the original vertices of \( S \). The following is a description of the system of labeling the vertices.

**Lemma 18.10.2** Let \( [x_0, \ldots, x_n] \) be an \( n \) simplex with \( \{x_k - x_0\}_{k=1}^n \) independent, and let the first \( n + 1 \) primes be \( p_0, p_1, \ldots, p_n \). Label \( x_k \) as \( p_k \) and consider a triangulation of this simplex. Labeling the vertices of this triangulation which occur on \( [x_{k_1}, \ldots, x_{k_s}] \) with any of \( p_{k_1}, \ldots, p_{k_s} \), beginning with all \( 1 \) simplices \( [x_{k_1}, x_{k_2}] \) and then \( 2 \) simplices and so forth, there are an odd number of simplices
18.10. BROUWER FIXED POINT THEOREM $\mathbb{R}^N$

\[ [y_{k_1}, \ldots, y_{k_s}] \] of the triangulation contained in \([x_{k_1}, \ldots, x_s]\) which have value \(p_{k_1}, \ldots, p_{k_s}\). This for \(s = 1, 2, \ldots, n\).

A combinatorial method

We now give a brief discussion of the system of labeling for Sperner’s lemma from the point of view of counting numbers of faces rather than obtaining them with an algorithm. Let \(p_0, \ldots, p_n\) be the first \(n + 1\) prime numbers. All vertices of a simplex \(S = [x_0, \ldots, x_n]\) having \([x_k - x_0]\) \(k=1\) independent will be labeled with one of these primes. In particular, the vertex \(x_k\) will be labeled as \(p_k\). The value of a simplex will be the product of its labels. Triangulate this \(S\). Consider a 1 simplex coming from the original simplex \([x_{k_1}, x_{k_2}]\), label one end as \(p_{k_1}\) and the other as \(p_{k_2}\). Then label all other vertices of this triangulation which occur on \([x_{k_1}, x_{k_2}]\) either \(p_{k_1}\) or \(p_{k_2}\). The assumption of linear independence assures that no other vertex of \(S\) can be in \([x_{k_1}, x_{k_2}]\) so there will be no inconsistency in the labeling. Then obviously there will be an odd number of simplices in this triangulation having value \(p_{k_1}p_{k_2}\), that is a \(p_{k_1}\) at one end and a \(p_{k_2}\) at the other. Suppose that the labeling has been done for all vertices of the triangulation which are on \([x_{j_1}, \ldots, x_{j_{k+1}}]\),

\[ \{x_{j_1}, \ldots, x_{j_{k+1}}\} \subseteq \{x_0, \ldots, x_n\} \]

any \(k\) simplex for \(k \leq n - 1\), and there is an odd number of simplices from the triangulation having value equal to \(\prod_{j=1}^{k+1} p_{j_i}\). Consider \(\tilde{S} \equiv [x_{j_1}, \ldots, x_{j_{k+1}}, x_{j_{k+2}}]\). Then by induction, there is an odd number of \(k\) simplices on the \(s^{th}\) face

\[ [x_{j_1}, \ldots, \tilde{x}_{j_s}, \ldots, x_{j_{k+1}}] \]

having value \(\prod_{j \neq s} p_{j_i}\). In particular the face \([x_{j_1}, \ldots, x_{j_{k+1}}, \tilde{x}_{j_{k+2}}]\) has an odd number of simplices with value \(\prod_{i=1}^{k+2} p_{j_i} := \tilde{P}_{k+1}\). We want to argue that some simplex in the triangulation which is contained in \(S\) has value \(\tilde{P}_{k+1} := \prod_{i=1}^{k+2} p_{j_i}\). Let \(Q\) be the number of \(k + 1\) simplices from the triangulation contained in \(\tilde{S}\) which have two faces with value \(\tilde{P}_k\) (A \(k + 1\) simplex has either 1 or 2 \(\tilde{P}_k\) faces.) and let \(R\) be the number of \(k + 1\) simplices from the triangulation contained in \(S\) which have exactly one \(\tilde{P}_k\) face. These are the ones we want because they have value \(\tilde{P}_{k+1}\). Thus the number of faces having value \(\tilde{P}_k\) which is described here is \(2Q + R\). All interior \(\tilde{P}_k\) faces being counted twice by this number. Now we count the total number of \(\tilde{P}_k\) faces another way. There are \(P\) of them on the face \([x_{j_1}, \ldots, x_{j_{k+1}}, \tilde{x}_{j_{k+2}}]\) and by induction, \(P\) is odd. Then there are \(O\) of them which are not on this face. These faces got counted twice. Therefore,

\[ 2Q + R = P + 2O \]

and so, since \(P\) is odd, so is \(R\). Thus there is an odd number of \(\tilde{P}_{k+1}\) simplices in \(\tilde{S}\).

We refer to this procedure of labeling as Sperner’s lemma. The system of labeling is well defined thanks to the assumption that \(\{x_k - x_0\}_{k=1}^n\) is independent which implies that \(\{x_k - x_i\}_{k \neq i}\) is also linearly independent. Thus there can be no ambiguity in the labeling of vertices on any “face”, the convex hull of some of the original vertices of \(S\). Sperner’s lemma is now a consequence of this discussion.

18.10.3 The Brouwer Fixed Point Theorem

\(S = [x_0, \ldots, x_n]\) is a simplex in \(\mathbb{R}^n\). Assume \(\{x_i - x_0\}_{i=1}^n\) are linearly independent. Thus a typical point of \(S\) is of the form

\[ \sum_{i=0}^n t_i x_i \]

where the \(t_i\) are uniquely determined and the map \(x \to t\) is continuous from \(S\) to the compact set \(\{t \in \mathbb{R}^{n+1} : \sum t_i = 1, t_i \geq 0\}\).
To see this, suppose \( x^k \to x \) in \( S \). Let \( x^k = \sum_{i=0}^n t_i^k x_i \) with \( x \) defined similarly with \( t_i^k \) replaced with \( t_i \), \( x = \sum_{i=0}^n t_i x_i \). Then

\[
x^k - x_0 = \sum_{i=0}^n t_i^k x_i - \sum_{i=0}^n t_i x_i = \sum_{i=1}^n t_i^k (x_i - x_0)
\]

Thus

\[
x^k - x_0 = \sum_{i=1}^n t_i^k (x_i - x_0), \quad x - x_0 = \sum_{i=1}^n t_i (x_i - x_0)
\]

Say \( t_i^k \) fails to converge to \( t_i \) for all \( i \geq 1 \). Then there exists a subsequence, still denoted with superscript \( k \) such that for each \( i = 1, \cdots, n \), it follows that \( t_i^k \to s_i \) where \( s_i \geq 0 \) and some \( s_i \neq t_i \). But then, taking a limit, it follows that

\[
x - x_0 = \sum_{i=1}^n s_i (x_i - x_0) = \sum_{i=1}^n t_i (x_i - x_0)
\]

which contradicts independence of the \( x_i - x_0 \). It follows that for all \( i \geq 1 \), \( t_i^k \to t_i \). Since they all sum to 1, this implies that also \( t_0^k \to t_0 \). Thus the claim about continuity is verified.

Let \( f : S \to S \) be continuous. When doing \( f \) to a point \( x \), one obtains another point of \( S \) denoted as \( \sum_{i=0}^n \varphi_i x_i \). Thus in this argument the scalars \( s_i \) will be the components after doing \( f \) to a point of \( S \) denoted as \( \sum_{i=0}^n t_i x_i \).

Consider a triangulation of \( S \) such that all simplices in the triangulation have diameter less than \( \varepsilon \). The vertices of the simplices in this triangulation will be labeled from \( p_0, \cdots, p_n \) the first \( n+1 \) prime numbers. If \( [y_0, \cdots, y_n] \) is one of these simplices in the triangulation, each vertex is of the form \( \sum_{i=0}^n t_i x_i \) where \( t_i \geq 0 \) and \( \sum_{i=0}^n t_i = 1 \). Let \( y_i \) be one of these vertices, \( y_i = \sum_{i=0}^n t_i x_i \). Define \( r_j = s_j/t_j \) if \( t_j \neq 0 \) and \( r_j = 0 \) if \( t_j = 0 \). Then \( p(y_i) \) will be the label placed on \( y_i \). To determine this label, let \( r_k \) be the smallest of these ratios. Then the label placed on \( y_i \) will be \( p_k \) where \( r_k \) is the smallest of all these extended nonnegative real numbers just described. If there is duplication, pick \( p_k \) where \( k \) is smallest.

Note that for the vertices which are on \([x_{i_1}, \cdots, x_{i_m}]\), these will be labeled from the list \( \{p_{i_1}, \cdots, p_{i_m}\} \) because \( t_k = 0 \) for each of these and so \( r_k = \infty \) unless \( k \in \{i_1, \cdots, i_m\} \). In particular, this scheme labels \( x_i \) as \( p_i \).

By the Sperner's lemma procedure described above, there are an odd number of simplices having value \( \prod_{j \neq k} p_j \) on the \( k \)th face and an odd number of simplices in the triangulation of \( S \) for which the product of the labels on their vertices, referred to here as its value, equals \( p_0 p_1 \cdots p_n = P_n \). Thus if \( [y_0, \cdots, y_n] \) is one of these simplices, and \( p(y_i) \) is the label for \( y_i \),

\[
\prod_{i=0}^n p(y_i) = \prod_{i=0}^n p_i \equiv P_n
\]

What is \( r_k \), the smallest of those ratios in determining a label? Could it be larger than 1? \( r_k \) is certainly finite because at least some \( t_j \neq 0 \) since they sum to 1. Thus, if \( r_k > 1 \), you would have \( s_k > t_k \). The \( s_j \) sum to 1 and so some \( s_j < t_j \) since otherwise, the sum of the \( t_j \) equalling 1 would require the sum of the \( s_j \) to be larger than 1. Hence \( r_k \) was not really the smallest after all and so \( r_k \leq 1 \). Hence \( s_k \leq t_k \).

Let \( S = \{S_1, \cdots, S_m\} \) denote those simplices whose value is \( P_n \). In other words, if \( \{y_0, \cdots, y_n\} \) are the vertices of one of these simplices in \( S \), and

\[
y_k = \sum_{i=0}^n t_i^k x_i
\]

\( r_k \leq r_j \) for all \( j \neq k \) and \( \{k_0, \cdots, k_n\} = \{0, \cdots, n\} \). Let \( b \) denote the barycenter of \( S_k = [y_0, \cdots, y_n] \),

\[
b = \sum_{i=0}^n \frac{1}{n+1} y_i
\]
18.11. INVARiANCE OF DOMAIN

Do the same system of labeling for each \( n + 1 \) simplex in a sequence of triangulations where the diameters of the simplices in the \( k^{th} \) triangulation is no more than \( 2^{-k} \). Thus each of these triangulations has a \( n + 1 \) simplex having diameter no more than \( 2^{-k} \) which has value \( P_n \). Let \( b_k \) be the barycenter of one of these \( n + 1 \) simplices having value \( P_n \). By compactness, there is a subsequence, still denoted with the index \( k \) such that \( b_k \to x \). This \( x \) is a fixed point.

Consider this last claim. \( x = \sum_{i=0}^{n} t_i x_i \) and after applying \( f \), the result is \( \sum_{i=0}^{n} s_i x_i \). Then \( b_k \) is the barycenter of some \( \sigma_k \) having diameter no more than \( 2^{-k} \) which has value \( P_n \). Say \( \sigma_k \) is a simplex having vertices \( \{y_{0}^{k}, \cdots, y_{n}^{k}\} \) and the value of \([y_{0}^{k}, \cdots, y_{n}^{k}]\) is \( P_n \). Thus also

\[
\lim_{k \to \infty} y_{i}^{k} = x.
\]

Re ordering these if necessary, we can assume that the label for \( y_{i}^{k} \) is \( p_i \) which implies that, as noted above, for each \( i = 0, \cdots, n \),

\[
\frac{s_i}{t_i} \leq 1, \quad s_i \leq t_i
\]

the \( i^{th} \) coordinate of \( f(y_{i}^{k}) \) with respect to the original vertices of \( S \) decreases and each \( i \) is represented for \( i = \{0, 1, \cdots, n\} \). As noted above,

\[
y_{i}^{k} \to x
\]

and so the \( i^{th} \) coordinate of \( y_{i}^{k}, t_{i}^{k} \) must converge to \( t_i \). Hence if the \( i^{th} \) coordinate of \( f(y_{i}^{k}) \) is denoted by \( s_{i}^{k}, \)

\[
s_{i}^{k} \leq t_{i}^{k}
\]

By continuity of \( f \), it follows that \( s_{i}^{k} \to s_{i} \). Thus the above inequality is preserved on taking \( k \to \infty \) and so

\[
0 \leq s_{i} \leq t_{i}
\]

this for each \( i \). But these \( s_{i} \) add to 1 as do the \( t_{i} \) and so in fact, \( s_{i} = t_{i} \) for each \( i \) and so \( f(x) = x \). This proves the following theorem which is the Brouwer fixed point theorem.

**Theorem 18.10.3** Let \( S \) be a simplex \([x_{0}, \cdots, x_{n}]\) such that \( \{x_{i} - x_{0}\}_{i=1}^{n} \) are independent. Also let \( f : S \to S \) be continuous. Then there exists \( x \in S \) such that \( f(x) = x \).

**Corollary 18.10.4** Let \( K \) be a closed convex bounded subset of \( \mathbb{R}^n \). Let \( f : K \to K \) be continuous. Then there exists \( x \in K \) such that \( f(x) = x \).

**Proof:** Let \( S \) be a large simplex containing \( K \) and let \( P \) be the projection map onto \( K \). See Problem [240](#) on Page 474 for the necessary properties of this projection map. Consider \( g(x) \equiv f(Px) \). Then \( g \) satisfies the necessary conditions for Theorem 18.10.3 and so there exists \( x \in S \) such that \( g(x) = x \). But this says \( x \in K \) and so \( g(x) = f(x) \). \( \blacksquare \)

The proof of this corollary is pretty significant. By a homework problem, a closed convex set is a retract of \( \mathbb{R}^n \). This is what it means when you say there is this continuous projection map which maps onto the closed convex set but does not change any point in the closed convex set. When you have a set \( A \) which is a subset of a set \( B \) which has the property that continuous functions \( f : B \to B \) have fixed points, and there is a continuous map \( P \) from \( B \) to \( A \) which leaves points of \( A \) unchanged, then it follows that \( A \) will have the same “fixed point property”. You can probably imagine all sorts of sets which are retracts of closed convex bounded sets.

### 18.11 Invariance Of Domain

As another application of the inverse function theorem is a simple proof of the important invariance of domain theorem which says that continuous and one to one functions defined on an open set in \( \mathbb{R}^n \) with values in \( \mathbb{R}^n \) take open sets to open sets.
Lemma 18.11.1 Let $f$ be continuous and map $\overline{B(p,r)} \subseteq \mathbb{R}^n$ to $\mathbb{R}^n$. Suppose that for all $x \in \overline{B(p,r)}$,
\[ |f(x) - x| < \varepsilon r \]
Then it follows that
\[ f\left(\overline{B(p,r)}\right) \supseteq B(p,(1-\varepsilon)r) \]

Proof: This is from the Brouwer fixed point theorem, Corollary 18.10.4. Consider for $y \in B(p,(1-\varepsilon)r)$
\[ h(x) = x - f(x) + y \]
Then $h$ is continuous and for $x \in \overline{B(p,r)}$,
\[ |h(x) - p| = |x - f(x) + y - p| < \varepsilon r + |y - p| < \varepsilon r + (1 - \varepsilon)r = r \]
Hence $h : \overline{B(p,r)} \rightarrow \overline{B(p,r)}$ and so it has a fixed point $x$ by Corollary 18.10.4. Thus
\[ x - f(x) + y = x \]
so $f(x) = y$. ■

The notation $\|f\|_K$ means $\sup_{x \in K}|f(x)|$.

Lemma 18.11.2 Let $K$ be a compact set in $\mathbb{R}^n$ and let $g : K \rightarrow \mathbb{R}^n, z \in K$ is fixed. Let $\delta > 0$. Then there exists a polynomial $q$ (each component a polynomial) such that
\[ \|q - g\|_K < \delta, \quad q(z) = g(z), \quad Dq(z)^{-1} \text{ exists} \]

Proof: By the Weierstrass approximation theorem, Theorem 18.2.7, there exists a polynomial $\tilde{q}$ such that
\[ \|\tilde{q} - g\|_K < \frac{\delta}{3} \]
Then define for $y \in K$
\[ q(y) = \tilde{q}(y) + g(z) - \tilde{q}(z) \]
Then
\[ q(z) = \tilde{q}(z) + g(z) - \tilde{q}(z) = g(z) \]
Also
\[ |q(y) - g(y)| \leq \|\tilde{q}(y) + g(z) - \tilde{q}(z) - g(y)\| \]
\[ \leq \|\tilde{q}(y) - g(y)\| + |g(z) - \tilde{q}(z)| < \frac{2\delta}{3} \]
and so since $y$ was arbitrary,
\[ \|q - g\|_K \leq \frac{2\delta}{3} < \delta \]
If $Dq(z)^{-1}$ exists, then this is what is wanted. If not, let
\[ 0 < \eta < \{|\lambda| : \lambda \text{ is an eigenvalue of } Dq(z), \lambda \neq 0\} \]
Then if $\eta$ is small enough, $q(y)$ could be replaced with $q(y) + \eta(y - z)$ and the above inequality would be preserved along with $q(z) = g(z)$ but now $Dq(z)$ would have no zero eigenvalues and would therefore be invertible. Simply use the modified $q$. ■

Lemma 18.11.3 Let $f : \overline{B(p,r)} \rightarrow \mathbb{R}^n$ where the ball is also in $\mathbb{R}^n$. Let $f$ be one to one, $f$ continuous. Then there exists $\delta > 0$ such that
\[ f\left(\overline{B(p,r)}\right) \supseteq B(f(p),\delta) \].
Proof: Since \( f \left( B(p, r) \right) \) is compact, it follows that \( f^{-1} : f \left( B(p, r) \right) \to B(p, r) \) is continuous. By Lemma 18.11.2, there exists a polynomial \( q : f \left( B(p, r) \right) \to \mathbb{R}^n \) such that

\[
\| q - f^{-1} \|_{f(B(p, r))} < \varepsilon r, \quad \varepsilon < 1, \quad Dq(f(p))^{-1} \text{ exists, and } q(f(p)) = p
\]

From the first inequality in the above,

\[
|q(f(x)) - x| = |q(f(x)) - f^{-1}(f(x))| \leq \| q - f^{-1} \|_{f(B(p, r))} < \varepsilon r
\]

By Lemma 18.11.4,

\[
q \circ f \left( B(p, r) \right) \supseteq B(p, (1 - \varepsilon) r)
\]

By the Inverse function theorem, there is an open set containing \( f(p) \) denoted as \( W \) such that on \( W, q \) is one to one and it and its inverse, defined on an open set \( V = q(W) \) both map open sets to open sets. By the construction, \( p \in V \) and so if \( \eta \) is small enough, it follows that \( B(p, \eta) \subseteq B(p, (1 - \varepsilon) r) \cap V \). Thus

\[
q \circ f \left( B(p, r) \right) \supseteq B(p, \eta)
\]

and \( q, q^{-1} \) both map open sets to open sets. Thus \( q^{-1}(B(p, \eta)) \) is an open set containing \( f(p) \). Hence if \( \delta \) is small enough,

\[
f \left( B(p, r) \right) \supseteq B(f(p), \delta) \]

With this lemma, the invariance of domain theorem comes right away. This remarkable theorem states that if \( f : U \to \mathbb{R}^n \) for \( U \) an open set in \( \mathbb{R}^n \) and if \( f \) is one to one and continuous, then \( f(U) \) is also an open set in \( \mathbb{R}^n \).

Theorem 18.11.4 Let \( U \) be an open set in \( \mathbb{R}^n \) and let \( f : U \to \mathbb{R}^n \) be one to one and continuous. Then \( f(U) \) is also an open subset in \( \mathbb{R}^n \).

Proof: It suffices to show that if \( p \in U \) then \( f(p) \) is an interior point of \( f(U) \). Let \( B(p, r) \subseteq U \). By Lemma 18.11.3, \( f(U) \supseteq f \left( B(p, r) \right) \supseteq B(f(p), \delta) \) so \( f(p) \) is indeed an interior point of \( f(U) \).

18.12 Tensor Products

This is on the tensor product space. It is a way to consider multilinear forms. There is a space called the tensor product space which has a norm. Then when you have a multilinear form with values in some space \( V \), it can be represented as a linear map from this tensor product space to \( V \). Some people like this notation because the map just mentioned is linear. To do this right, you need some material which is equivalent to the axiom of choice. This is in an appendix.

Definition 18.12.1 Denote by \( B(X \times Y, K) \) the bilinear forms having values in \( K \) a vector space. Define \( \otimes : X \times Y \to \mathcal{L}(B(X \times Y, K), K) \)

\[
x \otimes y(A) = A(x, y)
\]

where \( A \) is a bilinear form. Often \( K \) is a field, but it could be any vector space with no change.

Remark 18.12.2 \( \otimes \) is bilinear. That is

\[
(ax_1 + bx_2) \otimes y = a(x_1 \otimes y) + b(x_2 \otimes y)
\]

\[
x \otimes (ay_1 + by_2) = a(x \otimes y_1) + b(x \otimes y_2)
\]
This follows right away from the definition. Note that \( x \otimes y \in \mathcal{L}(B(X \times Y, K), K) \) because it is linear on \( B(X \times Y, K) \).

\[
x \otimes y (aA + bB) = aA (x, y) + bB (x, y) = a (x \otimes y) (A) + b (x \otimes y) (B)
\]

**Definition 18.12.3** Define \( X \otimes Y \) as the span of the \( x \otimes y \) in \( \mathcal{L}(B(X \times Y, K), K) \).

**Lemma 18.12.4** Let \( V \subseteq X \). That is, \( V \neq 0 \) is a subspace of \( X \). Let \( B_V \) be a Hamel basis for \( V \). Then there exists a Hamel basis for \( X, B_X \) which includes \( B_V \).

**Proof:** Consider \( (Y, B_Y) \) where \( Y \supseteq V \) and \( B_Y \supseteq B_V \). Call it \( \mathcal{F} \). Let \( (Y, B_Y) \subseteq \left( \tilde{Y}, B_{\tilde{Y}} \right) \) if and only if \( Y \subseteq \tilde{Y} \) and \( B_Y \subseteq B_{\tilde{Y}} \). Then let \( \mathcal{C} \) denote a maximal chain. Let \( \left( \tilde{X}, B_{\tilde{X}} \right) \) consist of the union of all \( Y \) in the chain along with the union of all \( B_Y \) in the chain. Then in fact this is something in \( \mathcal{F} \) thanks to this being a chain. Thus \( \tilde{X} = \text{span} \left( B_{\tilde{X}} \right) \). Then it must be the case that \( \tilde{X} = X \). If not, there exists \( y \notin \text{span} \left( B_{\tilde{X}} \right) \). But then \( \left( \text{span} \left( B_{\tilde{X}}, y \right), \{ B_{\tilde{X}} \}, \{ y \} \right) \) could be added to \( \mathcal{C} \) and obtain a strictly longer chain. \( \blacksquare \)

Next,

\[
x \otimes (ay + bz) (A) = A (x, ay + bz) = aA (x, y) + bA (x, z) = ax \otimes y (A) + bx \otimes z (A) = (ax \otimes y + bx \otimes z) (A)
\]

and so \( \otimes \) distributes across addition from the left and you can factor out scalars. Similarly, it will work the same way from the right.

**Lemma 18.12.5** Let \( B_X \) be a Hamel basis for \( X \). Then define \( L : X \rightarrow Y \) as follows. For each \( x \in B_X \), let \( Lx \equiv y_x \). Then for arbitrary \( z \), define \( Lz \) as follows. For \( z = \sum_{x \in B_X} c_x x \) where this denotes a finite sum consisting of the unique linear combination of basis vectors which equals \( x \),

\[
Lz = \sum_{x \in B_X} c_x y_x
\]

Then \( L \) is linear.

**Proof:** There is only one finite linear combination equal to \( z \) thanks to linear independence. Thus \( L \) is well defined. Why is it linear? Let \( z = \sum_{x \in B_X} c_x x, w = \sum_{x \in B_X} c_w x \). Then \( L (az + bw) = \)

\[
L \left( a \sum_{x \in B_X} c_x x + b \sum_{x \in B_X} c_w x \right) = a \sum_{x \in B_X} c_x y_x + b \sum_{x \in B_X} c_w y_x = aLz + bLw
\]

Thus \( L \) is linear. \( \blacksquare \)

**Proposition 18.12.6** Let \( X, Y \) be vector spaces and let \( E \) and \( F \) be linearly independent subsets of \( X, Y \) respectively. Then \( \{ e \otimes f : e \in E, f \in F \} \) is linearly independent in \( X \otimes Y \).

**Proof:** Let \( \kappa \in K \). Say \( \sum_{i=1}^{n} \lambda_i e_i \otimes f_i = 0 \) in \( X \otimes Y \). This means it sends everything in \( B(X \times Y, K) \) to 0. Let \( \psi \in F^* \), \( \psi (f_k) = 1 \) and \( \psi (f_i) = 0 \) for \( i \neq k \). Let \( \phi \in E^* \) be defined similarly. What you do is extend \( \{ e_i \} \) to a Hamel basis and then define \( \phi \) to equal 1 at \( e_k \) and \( \phi \) sends every other thing in the Hamel basis to 0. Then you look at \( \phi (x) \psi (y) \kappa = A (x, y) \). Then you have \( 0 = \sum_{i=1}^{n} \lambda_i e_i \otimes f_i (A) = \lambda e_k \phi (e_k) \psi (f_k) \kappa = \lambda \kappa \). Since \( \kappa \) is arbitrary it must be that \( \lambda \kappa = 0 \). Thus these are linearly independent. \( \blacksquare \)

**Proposition 18.12.7** Suppose \( u = \sum_{i=1}^{n} x_i \otimes y_i \) is in \( X \otimes Y \) and is a shortest representation. Then \( \{ x_i \}, \{ y_i \} \) are linearly independent. All such shortest representations have the same length. If \( \sum_i x_i \otimes y_i = 0 \) and \( \{ y_i \} \) are independent, then \( x_i = 0 \) for each \( i \). In particular, \( x \otimes y = 0 \iff x = 0 \) or \( y = 0 \).
Suppose delivers the given bilinear form. Thus contained in the span of the right side. Similarly, \( f \) can be done for each \( k \).

\[ \sum_{i=1}^n x_i \otimes y_i = \sum_{i=1}^{n-1} x_i \otimes y_i + x_n \otimes \sum_{j=1}^{n-1} a_j y_j \]

\[ = \sum_{i=1}^{n-1} x_i \otimes y_i + \sum_{i=1}^{n-1} a_i x_n \otimes y_i \]

\[ = \sum_{i=1}^{n-1} x_i \otimes y_i + a_i x_n \otimes y_i \]

\[ = \sum_{i=1}^{n-1} (x_i + a_i x_n) \otimes y_i \]

and so \( n \) was not smallest after all. Similarly \( \{x_i\} \) must be linearly independent. Now suppose that

\[ \sum_{i=1}^n x_i \otimes y_i = \sum_{j=1}^m u_j \otimes v_j \]

and both are of minimal length. Why is \( n = m \)? We know that \( \{x_i\}, \{y_i\}, \{u_j\}, \{v_j\} \) are independent. Let \( \psi_k(y_k) = 1, \psi \) sends all other vectors to 0. Then do both sides to \( A(x, y) \equiv \phi(x) \psi_k(y) \kappa \) where \( \kappa \in K \) is given, \( \phi \in X^* \) arbitrary.

\[ \sum_{i=1}^n x_i \otimes y_i (A) = \sum_{i=1}^n \phi(x_i) \psi_k(y_i) \kappa = \sum_{j=1}^m \phi(u_j) \psi(v_j) \kappa \]

Hence

\[ \phi \left( x_k - \sum_{j=1}^m u_j \psi(v_j) \right) \kappa = 0 \]

and this holds for any \( \phi \in X^* \) and for any \( \kappa \). Hence \( x_k \in \text{span} \{u_j\} \). Thus \( m \geq n \) since this can be done for each \( k \). \( \{x_1, \ldots, x_n\} \subseteq \text{span} (u_1, \ldots, u_m) \) and the left side is independent and is contained in the span of the right side. Similarly, \( \{u_1, \ldots, u_m\} \subseteq \text{span} (x_1, \ldots, x_n) \) and so \( m \leq n \). Thus \( m = n \).

Next suppose that \( \sum_{i=1}^n x_i \otimes y_i = 0 \) and the \( \{y_i\} \) are linearly independent. Then let \( \psi_k(y_k) = 1 \) and \( \psi_k \) sends the other vectors to 0. Then do the sum to \( A(x, y) = \phi(x) \psi(y) \kappa \) for \( \kappa \in K \). This yields \( \phi(x_k) \kappa = 0 \) for every \( \phi \). Hence \( x_k = 0 \). This is so for each \( k \). Similarly, if \( \sum_{i=1}^n x_i \otimes y_i = 0 \) and \( \{x_i\} \) independent, then each \( y_i = 0 \). \[ \square \]

The next theorem is very interesting. It is concerned with bilinear forms having values in \( V \) a vector space \( \psi : X \times Y \rightarrow V \). Roughly, it says there is a unique linear map from \( X \otimes Y \) which delivers the given bilinear form.

**Theorem 18.12.8** Suppose \( \psi \) is a bilinear map in \( B(X \times Y; V) \) where \( V \) is a vector space. Let \( \otimes : X \times Y \rightarrow X \otimes Y \) be given by \( \otimes(x, y) \equiv x \otimes y \). Then there exists a unique linear map \( \hat{\psi} \in \mathcal{L}(X \otimes Y; V) \) such that \( \hat{\psi} \circ \otimes = \psi \). In other words, the following diagram commutes.

\[ \begin{array}{c}
X \times Y \\
\otimes \downarrow \quad \psi \\
X \otimes Y \rightarrow V
\end{array} \]

That is, \( \hat{\psi}(x \otimes y) = \psi(x, y) \).
**Proof:** Let $E$ be a Hamel basis for $X$ and let $F$ be one for $Y$. Then by definition, $\{ e \otimes f : e \in E, f \in F \}$ spans $X \otimes Y$. To see this, suppose you have $x \otimes y$. Then $x = \sum_i a_i e_i, y = \sum_j b_j f_j$ and so

$$x \otimes y = \sum_i a_i e_i \otimes \sum_j b_j f_j = \sum_{i,j} a_i b_j e_i \otimes f_j$$

Also, it was shown above in Proposition 18.12.6 that this set is linearly independent. Therefore, it is a Hamel basis and you can define $\hat{\psi}$ as follows: $\hat{\psi} (e_i \otimes f_j) \equiv \psi (e_i, f_j)$ and extend $\hat{\psi}$ linearly to get the desired linear transformation $\hat{\psi}$. It is unique because its value on a Hamel basis is completely determined by the values of $\psi$. Thus $\hat{\psi}$ gives the right thing on $E \times F$. It gives the right thing on $X \times Y$ also. To see this, suppose you have $x \otimes y = \sum_{i,j} a_i b_j e_i \otimes f_j$ as above. Then

$$\hat{\psi} (x \otimes y) = \sum_{i,j} a_i b_j \hat{\psi} (e_i \otimes f_j) = \sum_{i,j} a_i b_j \psi (e_i, f_j) = \psi \left( \sum_i a_i e_i, \sum_j b_j f_j \right) = \psi (x, y) \blacksquare$$

You could define by analogy $X_1 \otimes X_2 \otimes \cdots \otimes X_p$ and all the above results would have analogs in this situation. Thus $x_1 \otimes x_2 \otimes \cdots \otimes x_p$ acts on a $p$ linear form $A$ as follows.

$$(x_1 \otimes x_2 \otimes \cdots \otimes x_p) A \equiv A (x_1, x_2, \cdots, x_p)$$

Then the above theorem has an obvious corollary.

**Corollary 18.12.9** Suppose $\psi$ is a $p$ linear map in $B(X_1 \times \cdots \times X_p; V)$ where $V$ is a vector space. Let $\otimes^p : X_1 \times \cdots \times X_p \rightarrow X_1 \otimes X_2 \otimes \cdots \otimes X_p$ be given by $\otimes^p (x_1, x_2, \cdots, x_p) \equiv x_1 \otimes x_2 \otimes \cdots \otimes x_p$. Then there exists a unique linear map $\hat{\psi} \in \mathcal{L}(X_1 \otimes X_2 \otimes \cdots \otimes X_p; V)$ such that $\hat{\psi} \circ \otimes^p = \psi$. In other words, the following diagram commutes.

$$\begin{array}{ccc}
X_1 \times \cdots \times X_p & \xrightarrow{\otimes^p} & X_1 \otimes X_2 \otimes \cdots \otimes X_p \\
\downarrow \psi & & \downarrow \hat{\psi} \\
& V \\
\end{array}$$

### 18.12.1 The Norm In Tensor Product Space

Let $X, Y$ be Banach spaces. Then we have the following definition.

**Definition 18.12.10** We define for $u \in X \otimes Y$

$$\pi (u) \equiv \inf \left\{ \sum_i \| x_i \| \| y_i \| : u = \sum_i x_i \otimes y_i \right\}$$

In this context, it is assumed that the elements of $X \otimes Y$ act on continuous bilinear forms. That is $A$ is bilinear and

$$|A(x, y)| \leq \|A\| \|x\| \|y\|$$

When we write $B(X \times Y; V)$ we mean the continuous ones. Here $V$ is a normed vector space and the underlying field for all is $F$.

**Proposition 18.12.11** $\pi (u)$ is well defined and is a norm. Also, $\pi (x \otimes y) = \|x\| \|y\|$.
Proof: It is obviously well defined. Let \( \lambda \in F \) where \( F \) is the field of interest. Assume \( \lambda \neq 0 \) since otherwise there is nothing to show. Then if \( u = \sum_i x_i \otimes y_i \),

\[
\pi (\lambda u) \leq \sum_i \| \lambda x_i \| \| y_i \| = |\lambda| \sum_i \| x_i \| \| y_i \|
\]

Then taking inf, it follows that

\[
\pi (\lambda u) \leq |\lambda| \pi (u)
\]

Then from what was just shown for arbitrary nonzero \( \lambda \),

\[
\pi (u) = \pi (\lambda^{-1} \lambda u) \leq \frac{1}{|\lambda|} \pi (\lambda u)
\]

and so \( |\lambda| \pi (u) \leq \pi (\lambda u) \). Hence \( \pi (\lambda u) = |\lambda| \pi (u) \).

Let \( u = \sum_i x_i \otimes y_i, v = \sum_j \hat{x}_j \otimes \hat{y}_j \) such that

\[
\pi (u) + \varepsilon > \sum_i \| x_i \| \| y_i \|, \quad \pi (v) + \varepsilon > \sum_j \| \hat{x}_j \| \| \hat{y}_j \|.
\]

Then

\[
\pi (u + v) \leq \sum_i \| x_i \| \| y_i \| + \sum_j \| \hat{x}_j \| \| \hat{y}_j \| \leq \pi (u) + \pi (v) + 2\varepsilon
\]

Since \( \varepsilon \) is arbitrary, \( \pi (u + v) \leq \pi (u) + \pi (v) \).

Now suppose \( \pi (u) = 0 \). Does it follow that \( u (A) = 0 \) for all \( A \in B (X \times Y, V) \)? Let \( A \in B (X \times Y; V) \) and let \( u = \sum_i x_i \otimes y_i \) such that

\[
\varepsilon/\|A\| = \pi (u) + \varepsilon/\|A\| > \sum_i \| x_i \| \| y_i \|
\]

Then

\[
u (A) = \sum_i A (x_i, y_i) \leq \sum_i \| A \| \| x_i \| \| y_i \| < (\varepsilon/\|A\|) \|A\| = \varepsilon
\]

and since \( \varepsilon \) is arbitrary, this requires \( u (A) = 0 \). Since \( A \) is arbitrary, this requires \( u = 0 \).

Next is the very interesting equality that \( \pi (x \otimes y) = \| x \| \| y \| \). It is obvious that

\[
\pi (x \otimes y) \leq \| x \| \| y \|
\]

because one way to write \( x \otimes y \) is \( x \otimes y \). Let \( \phi (x) = \| x \|, \psi (y) = \| y \| \) where \( \| \phi \|, \| \psi \| = 1 \). Here \( \phi, \psi \in X', Y' \) respectively. You get them from the Hahn Banach theorem. Then consider the continuous bilinear form \( A (x, y) = \phi (x) \psi (y) \). Say \( x \otimes y = \sum_i x_i \otimes y_i \). There is a linear map \( \psi \in L (X \otimes Y, V) \) such that \( \psi (x \otimes y) = A (x, y) \). You just specify this on all things of the form \( e \otimes f \) where \( e \in E \) a Hamel basis for \( X \) and \( f \in F \), a Hamel basis for \( Y \). Then it must hold for the linear span of these things which would yield the desired result. Hence, in particular,

\[
\| x \| \| y \| = \phi (x) \psi (y) = |A (x, y)| = |\psi (x \otimes y)| = \left| \sum_i \psi (x_i \otimes y_i) \right|
\]

\[
= \left| \sum_i A (x_i, y_i) \right| = \left| \sum_i \phi (x_i) \psi (y_i) \right| \leq \sum_i \| x_i \| \| y_i \|
\]

It follows that on taking inf of both sides over all such representations of \( x \otimes y \) that

\[
\| x \| \| y \| \leq \pi (x \otimes y).
\]

There is no difference if you replace \( X \otimes Y \) with \( X_1 \otimes X_2 \otimes \cdots \otimes X_p \). One modifies the definition as follows.
Definition 18.12.12 We define for \( u \in X_1 \otimes X_2 \otimes \cdots \otimes X_p \)

\[
\pi (u) = \inf \left\{ \sum_{i}^{p} \prod_{j=1}^{\infty} \left\| x^j_i \right\| : u = \sum_{i}^{p} x^1_i \otimes x^2_i \otimes \cdots \otimes x^p_i \right\}
\]

In this context, it is assumed that the elements of \( X_1 \otimes X_2 \otimes \cdots \otimes X_p \) act on continuous \( p \) linear forms. That is \( A \) is \( p \) linear and \( |A(x_1, \cdots, x_p)| \leq \| A \| \prod \| x_i \| \)

Corollary 18.12.13 \( \pi (u) \) is well defined and is a norm. Also,

\[
\pi \left( x^1 \otimes x^2 \otimes \cdots \otimes x^p \right) = \prod_{k} \left\| x^k \right\|.
\]

Recall Corollary 18.12.9

\[
\begin{align*}
X_1 \times \cdots \times X_p \\
\otimes^p \downarrow \psi \circ \psi \\
X_1 \otimes X_2 \otimes \cdots \otimes X_p \rightarrow \psi \circ \psi \rightarrow V
\end{align*}
\]

Is it the case that \( \hat{\psi} \) is continuous? Letting \( a \in X_1 \otimes X_2 \otimes \cdots \otimes X_p \), does it follow that \( \left\| \hat{\psi} (a) \right\|_V \leq \left\| a \right\|_{X_1 \otimes X_2 \otimes \cdots \otimes X_p} \).

Let \( a = \sum_i x^1_i \otimes x^2_i \otimes \cdots \otimes x^p_i \). Then

\[
\left\| \hat{\psi} (a) \right\|_V = \left\| \hat{\psi} \left( \sum_i x^1_i \otimes x^2_i \otimes \cdots \otimes x^p_i \right) \right\|_V = \left\| \sum_i \hat{\psi} \left( x^1_i \otimes x^2_i \otimes \cdots \otimes x^p_i \right) \right\|_V
\]

\[
= \left\| \sum_i A \left( x^1_i, \cdots, x^p_i \right) \right\|_V \leq \| A \| \sum_i \prod \| x^m_i \|
\]

Then taking the inf over all such representations of \( a \), it follows that \( \left\| \hat{\psi} (a) \right\| \leq \| A \| \| a \|_{X_1 \otimes X_2 \otimes \cdots \otimes X_p} \).

This proves the following theorem.

Theorem 18.12.14 Suppose \( A \) is a continuous \( p \) linear map in \( B \left( X_1 \times \cdots \times X_p; V \right) \) where \( V \) is a vector space. Let \( \otimes^p : X_1 \times \cdots \times X_p \rightarrow X_1 \otimes X_2 \otimes \cdots \otimes X_p \) be given by \( \otimes^p (x_1, x_2, \cdots, x_p) \equiv x_1 \otimes x_2 \otimes \cdots \otimes x_p \). Then there exists a unique linear map \( \hat{\psi} \in \mathcal{L} \left( X_1 \otimes X_2 \otimes \cdots \otimes X_p; V \right) \) such that \( \hat{\psi} \circ \otimes^p = \psi \). In other words, the following diagram commutes.

\[
\begin{align*}
X_1 \times \cdots \times X_p \\
\otimes^p \downarrow \psi \circ \psi \\
X_1 \otimes X_2 \otimes \cdots \otimes X_p \rightarrow \psi \circ \psi \rightarrow V
\end{align*}
\]

Also \( \hat{\psi} \) is a continuous linear mapping.

18.12.2 The Taylor Formula And Tensors

Let \( V \) be a Banach space and let \( P \) be a function defined on an open subset of \( V \) which has values in a Banach space \( W \). Recall the nature of the derivatives. \( DP (x) \in \mathcal{L} (V, W) \). Thus \( v \rightarrow DP (x) (v) \) is a differentiable map from \( V \) to \( W \). Then from the definition,

\[
\hat{v} \rightarrow D^2 P (x) (\hat{v})
\]
is again a differentiable function with values in $W$. $D^2 P (x) \in \mathcal{L}(V, \mathcal{L}(V, W))$. We can write it in the form

$$D^2 P (x) (v, \dot{v})$$

Of course it is linear in both variables from the definition. Similarly, we denote $P^j (x)$ as a $j$ linear form which is the $j^{th}$ derivative. In fact it is a symmetric $j$ linear form as long as it is continuous. To see this, consider the case where $j = 3$ and $P$ has values in $\mathbb{R}$. Then let

$$P (x + tu + sv + rw) = h (t, s, r)$$

By equality of mixed partial derivatives, $h_{txr} (0, 0, 0) = h_{rst} (0, 0, 0)$ etc. Thus $h_{txr} (t, s, r) =$

$$P^3 (x + tu + sv + rw) (u, v, w)$$

and so

$$P^k (x) (u, v, w) = P^3 (x) (v, u, w)$$

etc. If it has values in $W$ you just replace $P$ with $\phi (P)$ where $\phi \in V'$ and use this result to conclude that $\phi (P^3 (u, v, w)) = \phi (P^3 (v, u, w))$ etc. Then since the dual space separates points, the desired result is obtained. Similarly, $D^3 P (x) (v, u, w) = D^3 P (x) (u, v, w)$ etc.

Recall the following proof of Taylor’s formula:

$$h (t) = P (u + tv)$$

$$h (t) = h (0) + h' (0) t + \cdots + \frac{1}{k!} h^k (0) t + \frac{1}{(k + 1)!} h^{(k+1)} (s) |t| < |t|$$

Now

$$h' (t) = DP (u + tv) v,$$

$$h'' (t) = D^2 P (u + tv) (v) (v),$$

$$h''' (t) = D^3 P (u + tv) (v) (v) (v)$$

etc. Thus letting $t = 1$,

$$P (u + v) = P (u) + DP (u) v + \frac{1}{2!} D^2 P (u) (v, v) + \cdots + \frac{1}{k!} D^k P (u) (v, \cdots, v) + o \left( |v|^{k} \right)$$

Now you can use the representation theorem Theorem 18.12.14 to write this in the following form.

$$P (u + v) = P (u) + P^1 (u) v + \frac{1}{2!} D^2 (u) v \otimes v + \cdots + \frac{1}{k!} P^k (u) v \otimes \cdots \otimes v + o \left( |v|^{k} \right)$$

where $f^j (u) \in \mathcal{L}(V \otimes \cdots \otimes V, W)$. I think that the reason this is somewhat attractive is that the $P^k$ are linear operators. Recall $P^k (u) v \otimes v = DP^k (u) (v, \cdots, v)$.

18.13 Exercises

1. This problem was suggested to me by Matt Heiner. Earlier there was a problem in which two surfaces intersected at a point and this implied that in fact, they intersected in a smooth curve. Now suppose you have two spheres $x^2 + y^2 + z^2 = 1$ and $(x-2)^2 + y^2 + z^2 = 1$. These intersect at the single point $(1, 0, 0)$. Why does the implicit function theorem not imply that these surfaces intersect in a curve?

2. Maximize $2x + y$ subject to the condition that $\frac{x^2}{4} + \frac{y^2}{9} \leq 1$. **Hint:** You need to consider interior points and also the method of Lagrange multipliers for the points on the boundary of this ellipse.

3. Maximize $x + y$ subject to the condition that $x^2 + \frac{y^2}{9} + z^2 \leq 1$. **
4. Find the points on \( y^2x = 16 \) which are closest to \((0, 0)\).

5. Let \( f(x, y, z) = x^2 - 2xy + 2z^2 - 4z + 2 \). Identify all the points where \( Df = 0 \). Then determine whether they are local minima local maxima or saddle points.

6. Let \( f(x, y) = x^4 - 2x^2 + 2y^2 + 1 \). Identify all the points where \( Df = 0 \). Then determine whether they are local minima local maxima or saddle points.

7. Let \( f(x, y, z) = -x^4 + 2x^2 - y^2 - 2z^2 - 1 \). Identify all the points where \( Df = 0 \). Then determine whether they are local minima local maxima or saddle points.

8. Let \( f: V \to \mathbb{R} \) where \( V \) is a finite dimensional normed vector space. Suppose \( f \) is convex, which means
   \[
   f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)
   \]
   whenever \( t \in [0, 1] \). Suppose also that \( f \) is differentiable. Show then that for every \( x, y \in V \),
   \[
   (Df(x) - Df(y))(x - y) \geq 0.
   \]
   Thus convex functions have monotone derivatives.

9. Suppose \( B \) is an open ball in \( X \) and \( f: B \to Y \) is differentiable. Suppose also there exists \( L \in \mathcal{L}(X, Y) \) such that
   \[
   \|Df(x) - L\| < k
   \]
   for all \( x \in B \). Show that if \( x_1, x_2 \in B \),
   \[
   |f(x_1) - f(x_2) - L(x_1 - x_2)| \leq k|x_1 - x_2|.
   \]
   **Hint:** Consider \( Tx = f(x) - Lx \) and argue \( \|DT(x)\| < k \).

10. Let \( f: U \subseteq X \to Y \), \( Df(x) \) exists for all \( x \in U \), \( B(x_0, \delta) \subseteq U \), and there exists \( L \in \mathcal{L}(X, Y) \), such that \( L^{-1} \in \mathcal{L}(Y, X) \), and for all \( x \in B(x_0, \delta) \)
    \[
    \|Df(x) - L\| < \frac{r}{\|L^{-1}\|}, \quad r < 1.
    \]
    Show that there exists \( \varepsilon > 0 \) and an open subset of \( B(x_0, \delta), V \), such that \( f: V \to B(f(x_0), \varepsilon) \) is one to one and onto. Also \( Df^{-1}(y) \) exists for each \( y \in B(f(x_0), \varepsilon) \) and is given by the formula
    \[
    Df^{-1}(y) = \left[Df\left(f^{-1}(y)\right)\right]^{-1}.
    \]
    **Hint:** Let
    \[
    T_y(x) \equiv T(x, y) = x - L^{-1}(f(x) - y)
    \]
    for \( |y - f(x_0)| < \frac{(1-r)\delta}{2\|L^{-1}\|} \), consider \( \{T^n_y(x_0)\} \). This is a version of the inverse function theorem for \( f \) only differentiable, not \( C^1 \).

11. In the last assignment, you showed that if \( Df(x_0) \) is invertible, then locally the function \( f \) was one to one. However, this is a strictly local result! Let \( f: \mathbb{R}^2 \to \mathbb{R}^2 \) be given by
    \[
    f(x, y) = \begin{pmatrix} e^x \cos y \\ e^x \sin y \end{pmatrix}
    \]
    This clearly is not one to one because if you replace \( y \) with \( y + 2\pi \), you get the same value. Now verify that \( Df(x, y)^{-1} \) exists for all \((x, y)\).

12. Show every polynomial, \( \sum_{|\alpha| \leq k} d_\alpha x^\alpha \) is \( C^k \) for every \( k \).
13. Suppose \( U \subseteq \mathbb{R}^2 \) is an open set and \( f : U \to \mathbb{R}^3 \) is \( C^1 \). Suppose \( Df (s_0, t_0) \) has rank two and

\[
f(s_0, t_0) = \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix}.
\]

Show that for \((s, t)\) near \((s_0, t_0)\), the points \( f(s, t) \) may be realized in one of the following forms.

\[
\{(x, y, \phi(x, y)) : (x, y) \text{ near } (x_0, y_0)\},
\]

\[
\{(\phi(y, z), y, z) : (y, z) \text{ near } (y_0, z_0)\},
\]

or

\[
\{(x, \phi(x, z), z) : (x, z) \text{ near } (x_0, z_0)\}.
\]

This shows that parametrically defined surfaces can be obtained locally in a particularly simple form.

14. Minimize \( \sum_{j=1}^{n} x_j \) subject to the constraint \( \sum_{j=1}^{n} x_j^2 = a^2 \). Your answer should be some function of \( a \) which you may assume is a positive number.

15. A curve is formed from the intersection of the plane, \( 2x+3y+z = 3 \) and the cylinder \( x^2+y^2 = 4 \). Find the point on this curve which is closest to \((0,0,0)\).

16. A curve is formed from the intersection of the plane, \( 2x+3y+z = 3 \) and the sphere \( x^2+y^2+z^2 = 16 \). Find the point on this curve which is closest to \((0,0,0)\).

17. Let \( A = (A_{ij}) \) be an \( n \times n \) matrix which is symmetric. Thus \( A_{ij} = A_{ji} \) and recall \( (Ax)_i = A_{ij}x_j \) where as usual sum over the repeated index. Show \( \frac{\partial}{\partial x_i} (A_{ij}x_j, x_i) = 2A_{ij}x_j \). Show that when you use the method of Lagrange multipliers to maximize the function, \( A_{ij}x_j, x_i \) subject to the constraint, \( \sum_{j=1}^{n} x_j^2 = 1 \), the value of \( \lambda \) which corresponds to the maximum value of this functions is such that \( A_{ij}x_j = \lambda x_i \). Thus \( Ax = \lambda x \). Thus \( \lambda \) is an eigenvalue of the matrix, \( A \).

18. Let \( x_1, \ldots, x_5 \) be 5 positive numbers. Maximize their product subject to the constraint that

\[
x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 = 300.
\]

19. Let \( f(x_1, \ldots, x_n) = x_1^n x_2^{n-1} \cdots x_n^1 \). Then \( f \) achieves a maximum on the set,

\[
S \equiv \left\{ x \in \mathbb{R}^n : \sum_{i=1}^{n} ix_i = 1 \text{ and each } x_i \geq 0 \right\}.
\]

If \( x \in S \) is the point where this maximum is achieved, find \( x_1/x_n \).

20. Maximize \( \prod_{i=1}^{n} x_i^2 \) \( (\equiv x_1^2 \times x_2^2 \times x_3^2 \times \cdots \times x_n^2) \) subject to the constraint, \( \sum_{i=1}^{n} x_i^2 = r^2 \). Show the maximum is \( (r^2/n)^n \). Now show from this that

\[
\left( \prod_{i=1}^{n} x_i^2 \right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^{n} x_i^2
\]

and finally, conclude that if each number \( x_i \geq 0 \), then

\[
\left( \prod_{i=1}^{n} x_i \right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^{n} x_i
\]

and there exist values of the \( x_i \) for which equality holds. This says the “geometric mean” is always smaller than the arithmetic mean.
21. Show that there exists a smooth solution \( y = y(x) \) to the equation
\[
xe^y + ye^x = 0
\]
which is valid for \( x, y \) both near 0. Find \( y'(x) \) at a point \((x, y)\) near \((0, 0)\). Then find \( y''(x) \) for such \((x, y)\). Can you find an explicit formula for \( y(x) \)?

22. The next few problems involve invariance of domain. Suppose \( U \) is a nonempty open set in \( \mathbb{R}^n \), \( f : U \to \mathbb{R}^n \) is continuous, and suppose that for each \( x \in U \), there is a ball \( B_x \) containing \( x \) such that \( f \) is one to one on \( B_x \). That is, \( f \) is locally one to one. Show that \( f(U) \) is open.

23. ↑ In the situation of the above problem, suppose \( f : \mathbb{R}^n \to \mathbb{R}^n \) is locally one to one. Also suppose that \( \lim_{|x| \to \infty} |f(x)| = \infty \). Show that it follows that \( f(\mathbb{R}^n) = \mathbb{R}^n \). That is, \( f \) is onto. Show that this would not be true if \( f \) is only defined on a proper open set. Also show that this would not be true if the condition \( \lim_{|x| \to \infty} |f(x)| = \infty \) does not hold. Hint: You might show that \( f(\mathbb{R}^n) \) is both open and closed and then use connectedness. To get an example in the second case, you might think of \( e^{x+iy} \). It does not include \( 0 + i0 \). Why not?

24. ↑ Show that if \( f : \mathbb{R}^n \to \mathbb{R}^n \) is \( C^1 \) and if \( Df(x) \) exists and is invertible for all \( x \in \mathbb{R}^n \), then \( f \) is locally one to one. Thus, from the above problem, if \( \lim_{|x| \to \infty} |f(x)| = \infty \), then \( f \) is also onto. Now consider \( f : \mathbb{R}^2 \to \mathbb{R}^2 \) given by
\[
f(x, y) = \begin{pmatrix} e^x \cos y \\ e^x \sin y \end{pmatrix}
\]
Show that this does not map onto \( \mathbb{R}^2 \). In fact, it fails to hit \((0, 0)\), but \( Df(x, y) \) is invertible for all \((x, y)\). Show why it fails to satisfy the limit condition.
Chapter 19

Abstract Measures And Measurable Functions

The Lebesgue integral is much better than the Riemann integral. This has been known for over 100 years. It is much easier to generalize to many dimensions and it is much easier to use in applications. That is why I am going to use it rather than struggle with an inferior integral. It is also this integral which is most important in probability. However, this integral is more abstract. This chapter will develop the abstract machinery necessary for this integral.

The next definition describes what is meant by a $\sigma$ algebra. This is the fundamental object which is studied in probability theory. The events come from a $\sigma$ algebra of sets. Recall that $\mathcal{P}(\Omega)$ is the set of all subsets of the given set $\Omega$. It may also be denoted by $2^\Omega$ but I won’t refer to it this way.

Definition 19.0.1 $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, the set of all subsets of $\Omega$, is called a $\sigma$ algebra if it contains $\emptyset, \Omega$, and is closed with respect to countable unions and complements. That is, if $\{A_n\}_{n=1}^{\infty}$ is countable and each $A_n \in \mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ also and if $A \in \mathcal{F}$, then $\Omega \setminus A \in \mathcal{F}$. It is clear that any intersection of $\sigma$ algebras is a $\sigma$ algebra. If $\mathcal{K} \subseteq \mathcal{P}(\Omega)$, $\sigma(\mathcal{K})$ is the smallest $\sigma$ algebra which contains $\mathcal{K}$.

If $\mathcal{F}$ is a $\sigma$ algebra, then it is also closed with respect to countable intersections. Here is why. Let $\{F_k\}_{k=1}^{\infty} \subseteq \mathcal{F}$. Then $(\bigcap_k F_k)^C = \bigcup_k F_k^C \in \mathcal{F}$ and so

$$\bigcap_k F_k = \left((\bigcap_k F_k)^C\right)^C = \left(\bigcup_k F_k^C\right)^C \in \mathcal{F}.$$  

Example 19.0.2 You could consider $\mathbb{N}$ and for your $\sigma$ algebra, you could have $\mathcal{P}(\mathbb{N})$. This satisfies all the necessary requirements. Note that in fact, $\mathcal{P}(S)$ works for any $S$. However, useful examples are not typically the set of all subsets.

19.1 Simple Functions And Measurable Functions

Recall that a $\sigma$ algebra is a collection of subsets of a set $\Omega$ which includes $\emptyset, \Omega$, and is closed with respect to countable unions and complements.

Definition 19.1.1 Let $(\Omega, \mathcal{F})$ be a measurable space, one for which $\mathcal{F}$ is a $\sigma$ algebra contained in $\mathcal{P}(\Omega)$. Let $f : \Omega \to X$ where $X$ is a metric space. Then $f$ is measurable means that $f^{-1}(U) \in \mathcal{F}$ whenever $U$ is open.

It is important to have a theorem about pointwise limits of measurable functions, those with the property that inverse images of open sets are measurable. The following is a fairly general such theorem which holds in the situations to be considered in these notes.
Theorem 19.1.2 Let \( \{f_n\} \) be a sequence of measurable functions mapping \( \Omega \) to \((X, d)\) where \((X, d)\) is a metric space and \((\Omega, \mathcal{F})\) is a measurable space. Suppose also that \( f(\omega) = \lim_{n \to \infty} f_n(\omega) \) for all \( \omega \). Then \( f \) is also a measurable function.

**Proof:** It is required to show \( f^{-1}(U) \) is measurable for all \( U \) open. Let

\[
V_m \equiv \left\{ x \in U : \text{dist} (x, U^C) > \frac{1}{m} \right\}.
\]

Thus, since dist is continuous,

\[
V_m \subseteq \left\{ x \in U : \text{dist} (x, U^C) \geq \frac{1}{m} \right\}
\]

and \( V_m \subseteq \overline{V_m} \subseteq V_{m+1} \) and \( \cup_m V_m = U \). Then since \( V_m \) is open,

\[
f^{-1}(V_m) = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} f_k^{-1}(V_m)
\]

and so

\[
f^{-1}(U) = \bigcup_{m=1}^{\infty} f^{-1}(V_m) = \bigcup_{m=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} f_k^{-1}(V_m) \subseteq \bigcup_{m=1}^{\infty} f^{-1}(\overline{V_m}) = f^{-1}(U)
\]

which shows \( f^{-1}(U) \) is measurable. \( \square \)

An important example of a metric space is of course \( \mathbb{R}, \mathbb{C}, \mathbb{F}^n \), where \( \mathbb{F} \) is either \( \mathbb{R} \) or \( \mathbb{C} \) and so forth. However, it is also very convenient to consider the metric space \((-\infty, \infty]\), the real line with \( \infty \) tacked on at the end. This can be considered as a metric space in a very simple way.

\[
\rho(x, y) = |\arctan(x) - \arctan(y)|
\]

with the understanding that \( \arctan(\infty) \equiv \pi/2 \). It is easy to show that this metric restricted to \( \mathbb{R} \) gives the same open sets on \( \mathbb{R} \) as the usual metric given by \( d(x, y) = |x - y| \) but in addition, allows the inclusion of that ideal point out at the end of the real line denoted as \( \infty \). This is considered mainly because it makes the development of the theory easier. The open sets in \((-\infty, \infty]\) are described in the following lemma.

**Lemma 19.1.3** The open balls in \((-\infty, \infty]\) consist of sets of the form \((a, b)\) for \( a, b \) real numbers and \((a, \infty]\). This is a separable metric space.

**Proof:** If the center of the ball is a real number, then the ball will result in an interval \((a, b)\) where \( a, b \) are real numbers. If the center of the ball is \( \infty \), then the ball results in something of the form \((a, \infty]\). It is obvious that this is a separable metric space with the countable dense set being \( \mathbb{Q} \) since every ball contains a rational number. \( \square \)

If you kept both \(-\infty \) and \( \infty \) with the obvious generalization that \( \arctan(-\infty) \equiv -\pi/2 \), then the resulting metric space would be a complete separable metric space. However, it is not convenient to include \(-\infty \), so this won’t be done. The reason is that it will be desired to make sense of things like \( f + g \).

Then for functions which have values in \((-\infty, \infty]\) we have the following extremely useful description of what it means for a function to be measurable.

**Lemma 19.1.4** Let \( f : \Omega \to (-\infty, \infty]\) where \( \mathcal{F} \) is a \( \sigma \) algebra of subsets of \( \Omega \). Here \((-\infty, \infty]\) is the metric space just described with the metric given by

\[
\rho(x, y) = |\arctan(x) - \arctan(y)|.
\]

Then the following are equivalent.

\[
f^{-1}([d, \infty]) \in \mathcal{F} \text{ for all finite } d,
\]
19.1. SIMPLE FUNCTIONS AND MEASURABLE FUNCTIONS

\[ f^{-1}((-\infty, d)) \in \mathcal{F} \] for all finite \( d \),
\[ f^{-1}([d, \infty]) \in \mathcal{F} \] for all finite \( d \),
\[ f^{-1}((-\infty, d]) \in \mathcal{F} \] for all finite \( d \),
\[ f^{-1}((a, b)) \in \mathcal{F} \] for all \( a < b, -\infty < a < b < \infty \).

Any of these equivalent conditions is equivalent to the function being measurable.

**Proof:** First note that the first and the third are equivalent. To see this, observe
\[ f^{-1}([d, \infty]) = \bigcap_{n=1}^{\infty} f^{-1}((d - 1/n, \infty)), \]
and so if the first condition holds, then so does the third.
\[ f^{-1}((d, \infty]) = \bigcup_{n=1}^{\infty} f^{-1}([d + 1/n, \infty]), \]
and so if the third condition holds, so does the first.

Similarly, the second and fourth conditions are equivalent. Now
\[ f^{-1}((-\infty, d]) = (f^{-1}((d, \infty]))^C, \]
so the first and fourth conditions are equivalent. Thus the first four conditions are equivalent and if any of them hold, then for \(-\infty < a < b < \infty\),
\[ f^{-1}((a, b)) = f^{-1}((-\infty, b]) \cap f^{-1}((a, \infty]) \in \mathcal{F}. \]
Finally, if the last condition holds,
\[ f^{-1}([d, \infty]) = (\bigcup_{k=1}^{\infty} f^{-1}((-k + d, d]))^C \in \mathcal{F} \]
and so the third condition holds. Therefore, all five conditions are equivalent.

Since \((-\infty, \infty)\) is a separable metric space, it follows from Theorem 11.1.29 that every open set \( U \) is a countable union of open intervals \( U = \bigcup_k I_k \) where \( I_k \) is of the form \((a, b)\) or \((a, \infty]\) and, as just shown if any of the equivalent conditions holds, then \( f^{-1}(U) = \bigcup_k f^{-1}(I_k) \in \mathcal{F} \). Conversely, if \( f^{-1}(U) \in \mathcal{F} \) for any open set \( U \in (-\infty, \infty) \), then \( f^{-1}((a, b)) \in \mathcal{F} \) which is one of the equivalent conditions and so all the equivalent conditions hold. \( \blacksquare \)

There is a fundamental theorem about the relationship of simple functions to measurable functions given in the next theorem.

**Definition 19.1.5** Let \( E \in \mathcal{F} \) for \( \mathcal{F} \) a \( \sigma \) algebra. Then
\[ \chi_E(\omega) = \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases} \]
This is called the indicator function of the set \( E \). Let \( s : (\Omega, \mathcal{F}) \to \mathbb{R} \). Then \( s \) is a simple function if it is of the form
\[ s(\omega) = \sum_{i=1}^{n} c_i \chi_{E_i}(\omega) \]
where \( E_i \in \mathcal{F} \) and \( c_i \in \mathbb{R} \), the \( E_i \) being disjoint. Thus simple functions have finitely many values and are measurable. In the next theorem, it will also be assumed that each \( c_i \geq 0 \).

Each simple function is measurable. This is easily seen as follows. First of all, you can assume the \( c_i \) are distinct because if not, you could just replace those \( E_i \) which correspond to a single value with their union. Then if you have any open interval \((a, b)\),
\[ s^{-1}((a, b)) = \bigcup \{ E_i : c_i \in (a, b) \} \]
and this is measurable because it is the finite union of measurable sets.
Theorem 19.1.6 Let \( f \geq 0 \) be measurable. Then there exists a sequence of nonnegative simple functions \( \{s_n\} \) satisfying
\[
0 \leq s_n(\omega) \\
\cdots s_n(\omega) \leq s_{n+1}(\omega) \\
f(\omega) = \lim_{n \to \infty} s_n(\omega) \text{ for all } \omega \in \Omega.
\]
If \( f \) is bounded, the convergence is actually uniform. Conversely, if \( f \) is nonnegative and is the pointwise limit of such simple functions, then \( f \) is measurable.

**Proof:** Letting \( I = \{ \omega : f(\omega) = \infty \} \), define
\[
t_n(\omega) = \sum_{k=0}^{2^n} \frac{k}{n} \mathcal{X}_{f^{-1}\left(\left[\frac{k}{n}, \frac{k+1}{n}\right]\right)}(\omega) + 2^n \mathcal{X}_f(\omega).
\]
Then \( t_n(\omega) \leq f(\omega) \) for all \( \omega \) and \( \lim_{n \to \infty} t_n(\omega) = f(\omega) \) for all \( \omega \). This is because \( t_n(\omega) = 2^n \) for \( \omega \in I \) and if \( f(\omega) \in \left[0, \frac{2^n+1}{n}\right) \), then
\[
0 \leq f(\omega) - t_n(\omega) \leq \frac{1}{n}.
\]
Thus whenever \( \omega \notin I \), the above inequality will hold for all \( n \) large enough. Let \( s_1 = t_1, s_2 = \max (t_1, t_2), s_3 = \max (t_1, t_2, t_3), \ldots \).

Then the sequence \( \{s_n\} \) satisfies \[19.1\]. Also each \( s_n \) has finitely many values and is measurable. To see this, note that
\[
s_{n-1}^{-1}(\left(a, \infty\right]) = \cup_{k=1}^{n} t_k^{-1}(\left(a, \infty\right]) \in \mathcal{F}
\]
To verify the last claim, note that in this case the term \( 2^n \mathcal{X}_f(\omega) \) is not present and for \( n \) large enough, \( 2^n/n \) is larger than all values of \( f \). Therefore, for all \( n \) large enough, \[19.6\] holds for all \( \omega \). Thus the convergence is uniform.

The last claim follows right away from Theorem \[19.6\].

Although it is not needed here, there is a more general theorem which applies to measurable functions which have values in a separable metric space. In this context, a simple function is one which is of the form
\[
\sum_{k=1}^{m} x_k \mathcal{X}_{E_k}(\omega)
\]
where the \( E_k \) are disjoint measurable sets and the \( x_k \) are in \( X \). I am abusing notation somewhat by using a sum. You can’t add in a general metric space. The symbol means the function has value \( x_k \) on the set \( E_k \).

Theorem 19.1.7 Let \((\Omega, \mathcal{F})\) be a measurable space and let \( f : \Omega \to X \) where \((X, d)\) is a separable metric space. Then \( f \) is a measurable function if and only if there exists a sequence of simple functions, \( \{f_n\} \) such that for each \( \omega \in \Omega \) and \( n \in \mathbb{N} \),
\[
d(f_n(\omega), f(\omega)) \geq d(f_{n+1}(\omega), f(\omega))
\]
and
\[
\lim_{n \to \infty} d(f_n(\omega), f(\omega)) = 0.
\]

**Proof:** Let \( D = \{x_k\}_{k=1}^{\infty} \) be a countable dense subset of \( X \). First suppose \( f \) is measurable. Then since in a metric space every open set is the countable intersection of closed sets, it follows \( f^{-1}(\text{closed set}) \in \mathcal{F} \). Now let \( D_n = \{x_k\}_{k=1}^{n} \). Let
\[
A_1 = \left\{ \omega : d(x_1, f(\omega)) = \min_{k \leq n} d(x_k, f(\omega)) \right\}
\]
That is, \( A_1 \) are those \( \omega \) such that \( f(\omega) \) is approximated best out of \( D_n \) by \( x_1 \). Why is this a measurable set? It is because \( \omega \to d(x, f(\omega)) \) is a real valued measurable function, being the composition of a continuous function, \( y \to d(x, y) \) and a measurable function, \( \omega \to f(\omega) \). Next let

\[
A_2 \equiv \left\{ \omega \notin A_1 : d(x_2, f(\omega)) = \min_{k \leq n} d(x_k, f(\omega)) \right\}
\]

and continue in this manner obtaining disjoint measurable sets, \( \{A_k\}_{k=1}^n \) such that for \( \omega \in A_k \) the best approximation to \( f(\omega) \) from \( D_n \) is \( x_k \). Then

\[
f_n(\omega) = \sum_{k=1}^{n} x_k \mathcal{X}_{A_k}(\omega).
\]

Note

\[
\min_{k \leq n+1} d(x_k, f(\omega)) \leq \min_{k \leq n} d(x_k, f(\omega))
\]

and so this verifies \(\text{19.1.2}\). It remains to verify \(\text{19.1.3}\).

Let \( \varepsilon > 0 \) be given and pick \( \omega \in \Omega \). Then there exists \( x_n \in D \) such that \( d(x_n, f(\omega)) < \varepsilon \). It follows from the construction that \( d(f_n(\omega), f(\omega)) \leq d(x_n, f(\omega)) < \varepsilon \). This proves the first half.

Now suppose the existence of the sequence of simple functions as described above. Each \( f_n \) is a measurable function because \( f_n^{-1}(U) = \cup \{A_k : x_k \in U\} \). Therefore, the conclusion that \( f \) is measurable follows from Theorem \(\text{19.1.2}\) on Page \(\text{19.1.2}\). ■

### 19.2 Measures And Their Properties

First we define what is meant by a measure.

**Definition 19.2.1** Let \( (\Omega, \mathcal{F}) \) be a measurable space. Here \( \mathcal{F} \) is a \( \sigma \) algebra of sets of \( \Omega \). Then \( \mu : \mathcal{F} \to [0, \infty] \) is called a measure if whenever \( \{F_i\}_{i=1}^\infty \) is a sequence of disjoint sets of \( \mathcal{F} \), it follows that

\[
\mu(\cup_{i=1}^\infty F_i) = \sum_{i=1}^\infty \mu(F_i)
\]

Note that the series could equal \( \infty \). If \( \mu(\Omega) < \infty \), then \( \mu \) is called a finite measure. An important case is when \( \mu(\Omega) = 1 \) when it is called a probability measure.

Note that \( \mu(\emptyset) = \mu(\emptyset \cup \emptyset) = \mu(\emptyset) + \mu(\emptyset) \) and so \( \mu(\emptyset) = 0 \).

**Example 19.2.2** You could have \( \mathcal{P}(\mathbb{N}) = \mathcal{F} \) and you could define \( \mu(S) \) to be the number of elements of \( S \). This is called counting measure. It is left as an exercise to show that this is a measure.

**Example 19.2.3** Here is a pathological example. Let \( \Omega \) be uncountable and \( \mathcal{F} \) will be those sets which have the property that either the set is countable or its complement is countable. Let \( \mu(E) = 0 \) if \( E \) is countable and \( \mu(E) = 1 \) if \( E \) is uncountable. It is left as an exercise to show that this is a measure.

Of course the most important measure is Lebesgue measure which gives the “volume” of a subset of \( \mathbb{R}^n \). However, this requires a lot more work.

**Lemma 19.2.4** If \( \mu \) is a measure and \( F_i \in \mathcal{F} \), then \( \mu(\cup_{i=1}^\infty F_i) \leq \sum_{i=1}^\infty \mu(F_i) \). Also if \( F_n \in \mathcal{F} \) and \( F_n \subseteq F_{n+1} \) for all \( n \), then if \( F = \cup_n F_n \),

\[
\mu(F) = \lim_{n \to \infty} \mu(F_n)
\]

If \( F_n \supseteq F_{n+1} \) for all \( n \), then if \( \mu(F_1) < \infty \) and \( F = \cap_n F_n \), then

\[
\mu(F) = \lim_{n \to \infty} \mu(F_n)
\]
Proof: Let $G_1 = F_1$ and if $G_1, \ldots, G_n$ have been chosen disjoint, let

$$G_{n+1} = F_{n+1} \setminus \cup_{i=1}^n G_i$$

Thus the $G_i$ are disjoint. In addition, these are all measurable sets. Now

$$\mu(G_{n+1}) + \mu(F_{n+1} \cap (\cup_{i=1}^n G_i)) = \mu(F_{n+1})$$

and so $\mu(G_n) \leq \mu(F_n)$. Therefore,

$$\mu(\cup_{i=1}^\infty G_i) = \sum_i \mu(G_i) \leq \sum_i \mu(F_i).$$

Now consider the increasing sequence of $F_n \in \mathcal{F}$. If $F \subseteq G$ and these are sets of $\mathcal{F}$

$$\mu(G) = \mu(F) + \mu(G \setminus F)$$

so $\mu(G) \geq \mu(F)$. Also

$$F = \cup_{i=1}^\infty (F_{i+1} \setminus F_i) + F_1$$

Then

$$\mu(F) = \sum_{i=1}^\infty \mu(F_{i+1} \setminus F_i) + \mu(F_1)$$

Now $\mu(F_{i+1} \setminus F_i) + \mu(F_i) = \mu(F_{i+1})$. If any $\mu(F_i) = \infty$, there is nothing to prove. Assume then that these are all finite. Then

$$\mu(F_{i+1} \setminus F_i) = \mu(F_{i+1}) - \mu(F_i)$$

and so

$$\mu(F) = \sum_{i=1}^\infty \mu(F_{i+1}) - \mu(F_i) + \mu(F_1)$$

$$= \lim_{n \to \infty} \sum_{i=1}^n \mu(F_{i+1}) - \mu(F_i) + \mu(F_1) = \lim_{n \to \infty} \mu(F_{n+1})$$

Next suppose $\mu(F_1) < \infty$ and \{F_n\} is a decreasing sequence. Then

$$F_1 \setminus F_n$$

is increasing to $F_1 \setminus F$ and so by the first part,

$$\mu(F_1) - \mu(F) = \mu(F_1 \setminus F) = \lim_{n \to \infty} \mu(F_1 \setminus F_n) = \lim_{n \to \infty} (\mu(F_1) - \mu(F_n))$$

This is justified because $\mu(F_1 \setminus F_n) + \mu(F_n) = \mu(F_1)$ and all numbers are finite by assumption. Hence

$$\mu(F) = \lim_{n \to \infty} \mu(F_n).$$

19.3 Dynkin’s Lemma

Dynkin’s lemma is a very useful result. It is used quite a bit in books on probability. It will be used here to obtain $n$ dimensional Lebesgue measure, also to establish an important result on regularity in the next section.

Definition 19.3.1 Let $\Omega$ be a set and let $\mathcal{K}$ be a collection of subsets of $\Omega$. Then $\mathcal{K}$ is called a $\pi$ system if $\emptyset, \Omega \in \mathcal{K}$ and whenever $A, B \in \mathcal{K}$, it follows $A \cap B \in \mathcal{K}$. 
The following is the fundamental lemma which shows these \( \pi \) systems are useful. This is due to Dynkin.

**Lemma 19.3.2** Let \( K \) be a \( \pi \) system of subsets of \( \Omega \), a set. Also let \( G \) be a collection of subsets of \( \Omega \) which satisfies the following three properties.

1. \( K \subseteq G \)
2. If \( A \in G \), then \( A^C \in G \)
3. If \( \{ A_i \}_{i=1}^{\infty} \) is a sequence of disjoint sets from \( G \) then \( \bigcup_{i=1}^{\infty} A_i \in G \).

Then \( G \supseteq \sigma (K) \), where \( \sigma (K) \) is the smallest \( \sigma \) algebra which contains \( K \).

**Proof:** First note that if

\[ H \equiv \{ G : \text{ all assumptions hold} \} \]

then \( \cap H \) yields a collection of sets which also satisfies \( H \). Therefore, I will assume in the argument that \( G \) is the smallest collection of subsets of \( \Omega \) which satisfies \( H \). This will give the conclusion that for \( A \in K \) and \( B \in G \), \( A \cap B \in G \). This information will then be used to show that if \( A, B \in G \) then \( A \cap B \in G \). From this it will follow very easily that \( G \) is a \( \sigma \) algebra which will imply it contains \( \sigma (K) \). Now here are the details of the argument.

Since \( K \) is given to be a \( \pi \) system, \( K \subseteq G \). Property \( 3 \) is obvious because if \( \{ B_i \} \) is a sequence of disjoint sets in \( G \), then

\[ A \cap \bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A \cap B_i \in G \]

because \( A \cap B_i \in G \) and the property \( 2 \) of \( G \).

It remains to verify Property \( 2 \). So let \( B \in G \). I need to verify that \( B^C \in G \). In other words, I need to show that \( A \cap B^C \in G \). However,

\[ A \cap B^C = (A^C \cup (A \cap B))^C \in G \]

Here is why. Since \( B \in G \), \( A \cap B \in G \) and since \( A \in K \subseteq G \) it follows \( A^C \in G \) by assumption \( 3 \). It follows from assumption \( 2 \) the union of the disjoint sets, \( A^C \) and \( (A \cap B) \) is in \( G \) and then from \( 1 \) the complement of their union is in \( G \). Thus \( G \) satisfies \( H \) and this implies since \( G \) is the smallest such, that \( G \supseteq G \). However, \( G \) is constructed as a subset of \( G \). This proves that for every \( B \in G \) and \( A \in K \), \( A \cap B \in G \). Now pick \( B \in G \) and consider

\[ G_B \equiv \{ A \in G : A \cap B \in G \} \]

I just proved \( K \subseteq G_B \). The other arguments are identical to show \( G_B \) satisfies \( H \) and is therefore equal to \( G \). This shows that whenever \( A, B \in G \) it follows \( A \cap B \in G \).

This implies \( G \) is a \( \sigma \) algebra. To show this, all that is left is to verify \( G \) is closed under countable unions because then it follows \( G \) is a \( \sigma \) algebra. Let \( \{ A_i \} \subseteq G \). Then let \( A'_i = A_{i+1} \)

\[
A'_{n+1} = A_{n+1} \setminus \left( \bigcup_{i=1}^{n} A_i \right) \\
= A_{n+1} \cap \left( \bigcap_{i=1}^{n} A_i^C \right) \\
= \bigcap_{i=1}^{n} (A_{n+1} \cap A_i^C) \in G
\]

because finite intersections of sets of \( G \) are in \( G \). Since the \( A'_i \) are disjoint, it follows

\[ \bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} A'_i \in G \]

Therefore, \( G \supseteq \sigma (K) \).
19.4 Measures And Regularity

It is often the case that Ω has more going on than to simply be a set. In particular, it is often the case that Ω is some sort of topological space, often a metric space. In this case, it is usually if not always the case that the open sets will be in the σ algebra of measurable sets. This leads to the following definition.

**Definition 19.4.1** A Polish space is a complete separable metric space. For a Polish space E or more generally a metric space or even a general topological space, \( B(E) \) denotes the Borel sets of E. This is defined to be the smallest σ algebra which contains the open sets. Thus it contains all open sets and closed sets and compact sets and many others.

Don’t ever try to describe a generic Borel set. Always work with the definition that it is the smallest σ algebra containing the open sets. Attempts to give an explicit description of a “typical” Borel set tend to lead nowhere because there are so many things which can be done. You can take countable unions and complements and then countable intersections of what you get and then another countable union followed by complements and on and on. You just can’t get a good useable description in this way. However, it is easy to see that something like

\[
\left( \bigcap_{j=1}^{\infty} \bigcup_{i=1}^{\infty} E_j \right)^C
\]

is a Borel set if the \( E_j \) are. This is useful.

For example, \( \mathbb{R} \) is a Polish space as is any separable Banach space. Amazing things can be said about finite measures on the Borel sets of a Polish space. First the case of a finite measure on a metric space will be considered.

**Definition 19.4.2** A measure, \( \mu \) defined on \( B(E) \) will be called inner regular if for all \( F \in B(E) \),

\[
\mu(F) = \sup \{ \mu(K) : K \subseteq F \text{ and } K \text{ is closed} \}
\]

A measure, \( \mu \) defined on \( B(E) \) will be called outer regular if for all \( F \in B(E) \),

\[
\mu(F) = \inf \{ \mu(V) : V \supseteq F \text{ and } V \text{ is open} \}
\]

When a measure is both inner and outer regular, it is called regular. Actually, it is more useful and likely more standard to refer to \( \mu \) being inner regular as

\[
\mu(F) = \sup \{ \mu(K) : K \subseteq F \text{ and } K \text{ is compact} \}
\]

Thus the word “closed” is replaced with “compact”.

For finite measures, defined on the Borel sets of a metric space, the first definition of regularity is automatic. These are always outer and inner regular provided inner regularity refers to closed sets.

**Lemma 19.4.3** Let \( \mu \) be a finite measure defined on \( B(X) \) where \( X \) is a metric space. Then \( \mu \) is regular.

**Proof:** First note every open set is the countable union of closed sets and every closed set is the countable intersection of open sets. Here is why. Let \( V \) be an open set and let

\[
K_k \equiv \{ x \in V : \text{dist} (x, V^C) \geq 1/k \}.
\]

Then clearly the union of the \( K_k \) equals \( V \). Next, for \( K \) closed let

\[
V_k \equiv \{ x \in X : \text{dist} (x, K) < 1/k \}.
\]

Clearly the intersection of the \( V_k \) equals \( K \). Therefore, letting \( V \) denote an open set and \( K \) a closed set,

\[
\mu(V) = \sup \{ \mu(K) : K \subseteq V \text{ and } K \text{ is closed} \}
\]

\[
\mu(K) = \inf \{ \mu(V) : V \supseteq K \text{ and } V \text{ is open} \}.
\]
Also since $V$ is open and $K$ is closed,
\[
\mu(V) = \inf \{ \mu(U) : U \supseteq V \text{ and } V \text{ is open} \}
\]
\[
\mu(K) = \sup \{ \mu(L) : L \subseteq K \text{ and } L \text{ is closed} \}
\]
In words, $\mu$ is regular on open and closed sets. Let
\[
F = \{ F \in B(X) \text{ such that } \mu \text{ is regular on } F \}.
\]
Then $F$ contains the open sets and the closed sets.

Suppose $F \in F$. Then there exists $V \supseteq F$ with $\mu(V \setminus F) < \varepsilon$. It follows $V^C \subseteq F^C$ and
\[
\mu(F^C \setminus V^C) = \mu(V \setminus F) < \varepsilon.
\]
Thus $F^C$ is inner regular. Since $F \in F$, there exists $K \subseteq F$ where $K$ is closed and $\mu(F \setminus K) < \varepsilon$. Then also $K^C \supseteq F^C$ and
\[
\mu(K^C \setminus F^C) = \mu(F \setminus K) < \varepsilon.
\]
Thus if $F \in F$ so is $F^C$.

Suppose now that $\{ F_i \} \subseteq F$, the $F_i$ being disjoint. Is $\bigcup F_i \in F$? There exists $K_i \subseteq F_i$ such that $\mu(K_i) + \varepsilon/2^i > \mu(F_i)$. Then
\[
\mu(\bigcup_{i=1}^\infty F_i) = \sum_{i=1}^\infty \mu(F_i) \leq \varepsilon + \sum_{i=1}^\infty \mu(K_i)
\]
\[
< 2\varepsilon + \sum_{i=1}^N \mu(K_i) = 2\varepsilon + \mu((\bigcup_{i=1}^N K_i))
\]
provided $N$ is large enough. Thus it follows $\mu$ is inner regular on $\bigcup_{i=1}^\infty F_i$. Why is it outer regular? Let $V_i \supseteq F_i$ such that $\mu(F_i) + \varepsilon/2^i > \mu(V_i)$ and
\[
\mu(\bigcup_{i=1}^\infty F_i) = \sum_{i=1}^\infty \mu(F_i) > -\varepsilon + \sum_{i=1}^\infty \mu(V_i) \geq -\varepsilon + \mu((\bigcup_{i=1}^\infty V_i))
\]
which shows $\mu$ is outer regular on $\bigcup_{i=1}^\infty F_i$. It follows $F$ contains the $\pi$ system consisting of open sets and so by the Lemma on $\pi$ systems, Lemma 19.4.4, $F$ contains $\sigma(\tau)$ where $\tau$ is the set of open sets. Hence $F$ contains the Borel sets and is itself a subset of the Borel sets by definition. Therefore, $F = B(X)$. ■

One can say more if the metric space is complete and separable. In fact in this case the above definition of inner regularity can be shown to imply the usual one where the closed sets are replaced with compact sets.

**Lemma 19.4.4** Let $\mu$ be a finite measure on a $\sigma$ algebra containing $B(X)$, the Borel sets of $X$, a separable complete metric space. Then if $C$ is a closed set,
\[
\mu(C) = \sup \{ \mu(K) : K \subseteq C \text{ and } K \text{ is compact} \}
\]
It follows that for a finite measure on $B(X)$ where $X$ is a Polish space, $\mu$ is inner regular in the sense that for all $F \in B(X)$,
\[
\mu(F) = \sup \{ \mu(K) : K \subseteq F \text{ and } K \text{ is compact} \}
\]

**Proof:** Let $\{ a_k \}$ be a countable dense subset of $C$. Thus $\bigcup_{k=1}^\infty B\left( a_k, \frac{1}{n} \right) \supseteq C$. Therefore, there exists $m_n$ such that
\[
\mu\left( C \setminus \bigcup_{k=1}^{m_n} B\left( a_k, \frac{1}{n} \right) \right) \equiv \mu(C \setminus C_n) < \frac{\varepsilon}{2n}.
\]
Now let \( K = C \cap (\cap_{n=1}^{\infty} C_n) \). Then \( K \) is a subset of \( C_n \) for each \( n \) and so for each \( \varepsilon > 0 \) there exists an \( \varepsilon \) net for \( K \) since \( C_n \) has a \( 1/n \) net, namely \( a_1, \ldots, a_m \). Since \( K \) is closed, it is complete and so it is also compact since it is complete and totally bounded, Theorem 19.3. Now

\[
\mu (C \setminus K) = \mu (\cup_{n=1}^{\infty} (C \setminus C_n)) < \sum_{n=1}^{\infty} \frac{\varepsilon}{2^n} = \varepsilon.
\]

Thus \( \mu (C) \) can be approximated by \( \mu (K) \) for \( K \) a compact subset of \( C \). The last claim follows from Lemma 19.4.3. □

An important example is the case of a random vector and its distribution measure.

**Definition 19.4.5** A measurable function \( \mathbf{X} : (\Omega, \mathcal{F}, \mu) \to Z \) a metric space is called a random variable when \( \mu (\Omega) = 1 \). For such a random variable, one can define a distribution measure \( \lambda_{\mathbf{X}} \) on the Borel sets of \( Z \) as follows.

\[
\lambda_{\mathbf{X}} (G) \equiv \mu (\mathbf{X}^{-1} (G))
\]

This is a well defined measure on the Borel sets of \( Z \) because it makes sense for every \( G \) open and \( \mathcal{G} \equiv \{ G \subseteq Z : \mathbf{X}^{-1} (G) \in \mathcal{F} \} \) is a \( \sigma \) algebra which contains the open sets, hence the Borel sets. Such a random variable is also called a random vector when \( Z \) is a vector space.

**Corollary 19.4.6** Let \( \mathbf{X} \) be a random variable with values in a separable complete metric space, \( Z \). Then \( \lambda_{\mathbf{X}} \) is an inner and outer regular measure defined on \( \mathcal{B} (Z) \).

What if the measure \( \mu \) is defined on a Polish space but is not finite. Sometimes one can still get the assertion that \( \mu \) is regular. In every case of interest in this book, the measure will also be \( \sigma \) finite.

**Definition 19.4.7** Let \( (E, \mathcal{B} (E), \mu) \) be a measurable space with the measure \( \mu \). Then \( \mu \) is said to be \( \sigma \) finite if there is a sequence of disjoint Borel sets \( \{ B_i \}_{i=1}^{\infty} \) such that \( \cup_{i=1}^{\infty} B_i = E \) and \( \mu (B_i) < \infty \).

One such example of a complete metric space and a measure which is finite on compact sets is the following where the closures of balls are compact. Thus, this involves finite dimensional situations essentially.

**Corollary 19.4.8** Let \( \Omega \) be a complete separable metric space (Polish space). Let \( \mu \) be a measure on \( \mathcal{B} (\Omega) \) which has the property that \( \mu (B) < \infty \) for every ball \( B \). Then \( \mu \) must be regular.

**Proof:** Let \( \mu_K (E) \equiv \mu (K \cap E) \). Then this is a finite measure if \( K \) is contained in a ball and is therefore, regular.

Let

\[
A_n \equiv B (x_0, n) \setminus B (x_0, n - 1),
\]

\( x_0 \in \Omega \) and let

\[
B_n = B (x_0, n + 1) \setminus \overline{B (x_0, n - 2)}
\]

Thus the \( A_n \) are disjoint and have union equal to \( \Omega \), and the \( B_n \) are open sets having finite measure which contain the respective \( A_n \). (If \( x \) is a point, let \( n \) be the first such that \( x \in B (x_0, n) \).) Also, for \( E \subseteq A_n \),

\[
\mu (E) = \mu_{B_n} (E)
\]

By Lemma 19.4.3 each \( \mu_{B_n} \) is regular. Let \( E \) be any Borel set with \( l < \mu (E) \). Then for \( n \) large enough,

\[
l < \sum_{k=1}^{n} \mu (E \cap A_k) = \sum_{k=1}^{n} \mu_{B_k} (E \cap A_k)
\]

Choose \( r < 1 \) such that also

\[
l < r \sum_{k=1}^{n} \mu_{B_k} (E \cap A_k)
\]
There exists a compact set $K_k$ contained in $E \cap A_k$ such that
\[
\mu_B(K_k) > r \mu_B(E \cap A_k).
\]
Then letting $K$ be the union of these, $K \subseteq E$ and
\[
\mu(K) = \sum_{k=1}^{n} \mu(K_k) = \sum_{k=1}^{n} \mu_B(K_k) > r \sum_{k=1}^{n} \mu_B(E \cap A_k) > l
\]
Thus this is inner regular.

To show outer regular, it suffices to assume $\mu(E) < \infty$ since otherwise there is nothing to prove. There exists an open $V_n$ containing $E \cap A_n$ which is contained in $B_n$ such that
\[
\mu_B(E \cap A_n) + \varepsilon/2^n > \mu_B(V_n).
\]
Then let $V$ be the union of all these $V_n$.
\[
\mu(V \setminus E) = \mu(\bigcup_k V_k \setminus \bigcup_k (E \cap A_k)) \leq \sum_{k=1}^{\infty} \mu(V_k \setminus (E \cap A_k))
\]
\[
= \sum_{k=1}^{\infty} \mu_B(V_k \setminus (E \cap A_k)) < \sum_{k=1}^{\infty} \frac{\varepsilon}{2^k} = \varepsilon
\]
It follows that $\mu(E) + \varepsilon > \mu(V)$.

19.5 When Is A Measure A Borel Measure?

You have an outer measure defined on the power set of some metric space. How can you tell that the $\sigma$-algebra of measurable sets includes the Borel sets? This is what is discussed here. This is a very important idea because, from the above, you can then assert regularity of the measure if, for example it is finite on any ball.

**Definition 19.5.1** For two sets, $A, B$ in a metric space, we define
\[
\text{dist}(A, B) \equiv \inf \{d(x, y) : x \in A, y \in B\}.
\]

**Theorem 19.5.2** Let $\mu$ be an outer measure on the subsets of $(X, d)$, a metric space. If
\[
\mu(A \cup B) = \mu(A) + \mu(B)
\]
whenever $\text{dist}(A, B) > 0$, then the $\sigma$-algebra of measurable sets $\mathcal{S}$ contains the Borel sets.

**Proof:** It suffices to show that closed sets are in $\mathcal{S}$, the $\sigma$-algebra of measurable sets, because then the open sets are also in $\mathcal{S}$ and consequently $\mathcal{S}$ contains the Borel sets. Let $K$ be closed and let $\mathcal{S}$ be a subset of $\Omega$. Is $\mu(S) \geq \mu(S \cap K) + \mu(S \setminus K)$? It suffices to assume $\mu(S) < \infty$. Let
\[
K_n \equiv \{x : \text{dist}(x, K) \leq \frac{1}{n}\}
\]
By Lemma [11.1.14] on Page 239, $x \rightarrow \text{dist}(x, K)$ is continuous and so $K_n$ is closed. By the assumption of the theorem,
\[
\mu(S) \geq \mu((S \cap K) \cup (S \setminus K_n)) = \mu(S \cap K) + \mu(S \setminus K_n)
\]
(19.6)
since $S \cap K$ and $S \setminus K_n$ are a positive distance apart. Now
\[
\mu(S \setminus K_n) \leq \mu(S \setminus K) \leq \mu(S \setminus K_n) + \mu((K_n \setminus K) \cap S).
\]
(19.7)
if \( \lim_{n \to \infty} \mu((K_n \setminus K) \cap S) = 0 \) then the theorem will be proved because this limit along with (19.6) implies \( \lim_{n \to \infty} \mu(S \setminus K_n) = \mu(S \setminus K) \) and then taking a limit in (19.7), \( \mu(S) \geq \mu(S \cap K) + \mu(S \setminus K) \) as desired. Therefore, it suffices to establish this limit.

Since \( K \) is closed, a point, \( x \notin K \) must be at a positive distance from \( K \) and so

\[
K_n \setminus K = \bigcup_{k=n}^{\infty} K_k \setminus K_{k+1}.
\]

Therefore

\[
\mu(S \cap (K_n \setminus K)) \leq \sum_{k=n}^{\infty} \mu(S \cap (K_k \setminus K_{k+1})). \tag{19.8}
\]

If

\[
\sum_{k=1}^{\infty} \mu(S \cap (K_k \setminus K_{k+1})) < \infty, \tag{19.9}
\]

then \( \mu(S \cap (K_n \setminus K)) \to 0 \) because it is dominated by the tail of a convergent series so it suffices to show (19.7).

By the construction, the distance between any pair of sets, \( S \cap (K_k \setminus K_{k+1}) \) for different even values of \( k \) is positive and the distance between any pair of sets, \( S \cap (K_k \setminus K_{k+1}) \) for different odd values of \( k \) is positive. Therefore,

\[
\sum_{k \text{ even, } k \leq M} \mu(S \cap (K_k \setminus K_{k+1})) + \sum_{k \text{ odd, } k \leq M} \mu(S \cap (K_k \setminus K_{k+1})). \tag{19.10}
\]

By the construction, the distance between any pair of sets, \( S \cap (K_k \setminus K_{k+1}) \) for different even values of \( k \) is positive and the distance between any pair of sets, \( S \cap (K_k \setminus K_{k+1}) \) for different odd values of \( k \) is positive. Therefore,

\[
\sum_{k \text{ even, } k \leq M} \mu(S \cap (K_k \setminus K_{k+1})) + \sum_{k \text{ odd, } k \leq M} \mu(S \cap (K_k \setminus K_{k+1})) \leq \mu \left( \bigcup_{k \text{ even, } k \leq M} (S \cap (K_k \setminus K_{k+1})) \right) + \mu \left( \bigcup_{k \text{ odd, } k \leq M} (S \cap (K_k \setminus K_{k+1})) \right)
\]

\[
\leq \mu(S) + \mu(S) = 2\mu(S)
\]

and so for all \( M, \sum_{k=1}^{M} \mu(S \cap (K_k \setminus K_{k+1})) \leq 2\mu(S) \) showing (19.7).

19.6 Measures And Outer Measures

The above is all fine but is pretty abstract. Here is a simple example. Let \( \Omega = \mathbb{N} \) and let the \( \sigma \) algebra be the set of all subsets of \( \mathbb{N}, P(\mathbb{N}) \). Then let \( \mu(A) \equiv \text{the number of elements of } A \). This is known as counting measure. You can verify that this is an example of a measure and \( \sigma \) algebra. However, we really want more interesting examples than this. There is also something called an outer measure which is defined on the set of all subsets.

**Definition 19.6.1** Let \( \Omega \) be a nonempty set and let \( \lambda : P(\Omega) \to [0, \infty) \) satisfy the following:

1. \( \lambda(\emptyset) = 0 \)
2. If \( A \subseteq B \), then \( \lambda(A) \leq \lambda(B) \)
3. \( \lambda(\bigcup_{i=1}^{\infty} E_i) \leq \sum_{i=1}^{\infty} \lambda(E_i) \)

Then \( \lambda \) is called an outer measure.
Every measure determines an outer measure. For example, suppose that \( \mu \) is a measure on \( \mathcal{F} \) a \( \sigma \) algebra of subsets of \( \Omega \). Then define

\[
\hat{\mu} (S) \equiv \inf \{ \mu (E) : E \supseteq S, \ E \in \mathcal{F} \}
\]

This is easily seen to be an outer measure. Also, we have the following Proposition.

**Proposition 19.6.2** Let \( \mu \) be a measure as just described. Then \( \hat{\mu} \) as defined above, is an outer measure and also, if \( E \in \mathcal{F} \), then \( \hat{\mu} (E) = \mu (E) \).

**Proof:** The first two properties of an outer measure are obvious. What of the third? If any \( \hat{\mu} (E_i) = \infty \), then there is nothing to show so suppose each of these is finite. Let \( F_i \subseteq E_i \) such that \( F_i \in \mathcal{F} \) and \( \hat{\mu} (E_i) + \frac{\varepsilon}{2^i} > \mu (F_i) \). Then

\[
\hat{\mu} (\bigcup_{i=1}^{\infty} E_i) \leq \mu (\bigcup_{i=1}^{\infty} F_i) \leq \sum_{i=1}^{\infty} \mu (F_i)
\]

\[
< \sum_{i=1}^{\infty} \left( \hat{\mu} (E_i) + \frac{\varepsilon}{2^i} \right) = \sum_{i=1}^{\infty} \hat{\mu} (E_i) + \varepsilon
\]

Since \( \varepsilon \) is arbitrary, this establishes the third condition. Finally, if \( E \in \mathcal{F} \), then by definition, \( \hat{\mu} (E) \leq \mu (E) \) because \( E \supseteq E \). Also, \( \mu (E) \leq \mu (F) \) for all \( F \in \mathcal{F} \) such that \( F \supseteq E \). If follows that \( \mu (E) \) is a lower bound of all such \( \mu (F) \) and so \( \hat{\mu} (E) \geq \mu (E) \).

### 19.7 Exercises

1. Show carefully that if \( \mathcal{G} \) is a set whose elements are \( \sigma \) algebras which are subsets of \( \mathcal{P} (\Omega) \), then \( \cap \mathcal{G} \) is also a \( \sigma \) algebra. Now let \( \mathcal{G} \subseteq \mathcal{P} (\Omega) \) satisfy property \( P \) if \( \mathcal{G} \) is closed with respect to complements and countable disjoint unions as in Dynkin’s lemma, and contains \( \emptyset \) and \( \Omega \). If \( \mathcal{H} \) is any set whose elements are \( \mathcal{G} \) which satisfy property \( P \), then \( \cap \mathcal{H} \) also satisfies property \( P \).

2. The Borel sets of a metric space \( (X, d) \) are the sets in the smallest \( \sigma \) algebra which contains the open sets. These sets are denoted as \( \mathcal{B} (X) \). Thus \( \mathcal{B} (X) = \sigma (\text{open sets}) \) where \( \sigma (\mathcal{F}) \) simply means the smallest \( \sigma \) algebra which contains \( \mathcal{F} \). Show \( \mathcal{B} (\mathbb{R}^n) = \sigma (\mathcal{P}) \) where \( \mathcal{P} \) consists of the half open rectangles which are of the form \( \prod_{i=1}^{n} [a_i, b_i) \).

3. Recall that \( f : (\Omega, \mathcal{F}) \to X \) where \( X \) is a metric space is measurable means \( f^{-1} (\text{open}) \in \mathcal{F} \). Show that if \( E \) is any set in \( \mathcal{B} (X) \), then \( f^{-1} (E) \in \mathcal{F} \). Thus, inverse images of Borel sets are measurable. Next consider \( f : (\Omega, \mathcal{F}) \to X \) being measurable and \( g : X \to Y \) is Borel measurable, meaning that \( g^{-1} (\text{open}) \in \mathcal{B} (X) \). Explain why \( g \circ f \) is measurable. **Hint:** You know that \( (g \circ f)^{-1} (U) = f^{-1} (g^{-1} (U)) \). For your information, it does not work the other way around. That is, measurable composed with Borel measurable is not necessarily measurable. In fact examples exist which show that if \( g \) is measurable and \( f \) is continuous, then \( g \circ f \) may fail to be measurable. However, these things are not for this course.

4. If you have \( X_i \) is a metric space, let \( X = \prod_{i=1}^{n} X_i \) with the metric

\[
d (x, y) \equiv \max \{ d_i (x_i, y_i), i = 1, 2, \cdots , n \}
\]

You considered this in an earlier problem. Show that any set of the form

\[
\prod_{i=1}^{n} E_i, \ E_i \in \mathcal{B} (X_i)
\]

is a Borel set. That is, the product of Borel sets is Borel. **Hint:** You might consider the continuous functions \( \pi_i : \prod_{j=1}^{n} X_j \to X_i \) which are the projection maps. Thus \( \pi_i (x) \equiv x_i \). Then \( \pi_i^{-1} (E_i) \) would have to be Borel measurable whenever \( E_i \in \mathcal{B} (X_i) \). Explain why. You know \( \pi_i \) is continuous. Why would \( \pi_i^{-1} (\text{Borel}) \) be a Borel set? Then you might argue that \( \prod_{i=1}^{n} E_i = \cap_{i=1}^{n} \pi_i^{-1} (E_i) \).
5. You have two finite measures defined on $\mathcal{B} (X)$, $\mu, \nu$. Suppose these are equal on every open set. Show that these must be equal on every Borel set. **Hint:** You should use Dynkin’s lemma to show this very easily.

6. Show that $(\mathbb{N}, \mathcal{P} (\mathbb{N}), \mu)$ is a measure space where $\mu (S)$ equals the number of elements of $S$. You need to verify that if the sets $E_i$ are disjoint, then $\mu (\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu (E_i)$.

7. Let $\Omega$ be an uncountable set and let $\mathcal{F}$ denote those subsets of $\Omega$ that are countable. Show that this is a $\sigma$ algebra. Next define the following measure. $\mu (A) = 1$ if $A$ is uncountable and $\mu (A) = 0$ if $A$ is countable. Show that $\mu$ is a measure.

8. Let $\mu (E) = 1$ if $0 \in E$ and $\mu (E) = 0$ if $0 \notin E$. Show this is a measure on $\mathcal{P} (\mathbb{R})$.

9. Give an example of a measure $\mu$ and a measure space and a decreasing sequence of measurable sets $\{E_n\}$ such that $\lim_{n \to \infty} \mu (E_n) \neq \mu (\bigcap_{i=1}^{\infty} E_i)$.

10. If you have a finite measure $\mu$ on $\mathcal{B} (X)$ where $X$ is a metric space, and if $F \in \mathcal{B} (X)$, show that there exists a set $E, G$ such that $G$ is an open set and $E$ is a countable union of closed sets such that $E \subseteq F \subseteq G$ and $\mu (G \setminus E) = 0$.

11. You have a measure space $(\Omega, \mathcal{F}, P)$ where $P$ is a probability measure on $\mathcal{F}$. Then you also have a measurable function $X : \Omega \to Z$ where $Z$ is some metric space. Thus $X^{-1} (U) \in \mathcal{F}$ whenever $U$ is open. Now define a measure on $\mathcal{B} (Z)$ denoted by $\lambda_X$ and defined by $\lambda_X (E) = P (\{\omega : X (\omega) \in E\})$. Explain why this yields a well defined probability measure on $\mathcal{B} (Z)$ which is regular. This is called the distribution measure.

12. Let $K \subseteq V$ where $K$ is closed and $V$ is open. Consider the following function.

$$f (x) = \frac{\text{dist} \ (x, V^c)}{\text{dist} \ (x, K) + \text{dist} \ (x, V^c)}$$

Explain why this function is continuous, equals 0 off $V$ and equals 1 on $K$.

13. Let $(\Omega, \mathcal{F})$ be a measurable space and let $f : \Omega \to X$ be a measurable function. Then $\sigma (f)$ denotes the smallest $\sigma$ algebra such that $f$ is measurable with respect to this $\sigma$ algebra. Show that $\sigma (f) = \{f^{-1} (E) : E \in \mathcal{B} (X)\}$. More generally, you have a whole set of measurable functions $\mathcal{S}$ and $\sigma (\mathcal{S})$ denotes the smallest $\sigma$ algebra such that each function in $\mathcal{S}$ is measurable. If you have an increasing list $S_t$ for $t \in [0, \infty)$, then $\sigma (S_t)$ will be what is called a filtration. You have a $\sigma$ algebra for each $t \in [0, \infty)$ and as $t$ increases, these $\sigma$ algebras get larger. This is an essential part of the construction which is used to show that Wiener process is a martingale. In fact the whole subject of martingales has to do with filtrations.

14. There is a monumentally important theorem called the Borel Cantelli lemma. It says the following. If you have a measure space $(\Omega, \mathcal{F}, \mu)$ and if $\{E_i\} \subseteq \mathcal{F}$ is such that $\sum_{i=1}^{\infty} \mu (E_i) < \infty$, then there exists a set $N$ of measure 0 ($\mu (N) = 0$) such that if $\omega \notin N$, then $\omega$ is in only finitely many of the $E_i$. **Hint:** You might look at the set of all $\omega$ which are in infinitely many of the $E_i$. First explain why this set is of the form $\bigcap_{i=1}^{\infty} \cup_{k \geq n} E_k$.

15. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. A sequence of functions $\{f_n\}$ is said to converge in measure to a measurable function $f$ if and only if for each $\varepsilon > 0$,

$$\lim_{n \to \infty} \mu (\omega : |f_n (\omega) - f (\omega)| > \varepsilon) = 0$$

Show that if this happens, then there exists a set of measure $N$ such that if $\omega \notin N$, then

$$\lim_{n \to \infty} f_n (\omega) = f (\omega) .$$

Also show that $\lim_{n \to \infty} f_n (\omega) = f (\omega)$, then $f_n$ converges in measure to $f$. 

16. Let $X, Y$ be separable metric spaces. Then $X \times Y$ can also be considered as a metric space with the metric
\[ \rho((x, y), (\hat{x}, \hat{y})) \equiv \max(d_X(x, \hat{x}), d_Y(y, \hat{y})) \]
Verify this. Then show that if $\mathcal{K}$ consists of sets $A \times B$ where $A, B$ are Borel sets in $X$ and $Y$ respectively, then it follows that $\sigma(\mathcal{K}) = \mathcal{B}(X \times Y)$, the Borel sets from $X \times Y$. Extend to the Cartesian product $\prod_i X_i$ of finitely many separable metric spaces.

19.8 An Outer Measure On $\mathcal{P}(\mathbb{R})$

A measure on $\mathbb{R}$ is like length. I will present something more general than length because it is no trouble to do so and the generalization is useful in many areas of mathematics such as probability.

This outer measure will end up determining a measure known as the Lebesgue Stieltjes measure.

**Definition 19.8.1** The following definition is important.
\[ F(x+) \equiv \lim_{y \to x^+} F(y), \quad F(x-) = \lim_{y \to x^-} F(y) \]
Thus one of these is the limit from the left and the other is the limit from the right.

In probability, one often has $F(x) \geq 0$, $F$ is increasing, and $F(x+) = F(x)$. This is the case where $F$ is a probability distribution function. In this case, $F(x) \equiv P(X \leq x)$ where $X$ is a random variable. In this case, $\lim_{x \to \infty} F(x) = 1$ but we are considering more general functions than this including the simple example where $F(x) = x$. This last example will end up giving Lebesgue measure on $\mathbb{R}$.

**Definition 19.8.2** $\mathcal{P}(S)$ denotes the set of all subsets of $S$.

**Theorem 19.8.3** Let $F$ be an increasing function defined on $\mathbb{R}$. This will be called an integrator function. There exists a function $\mu : \mathcal{P}(\mathbb{R}) \to [0, \infty]$ which satisfies the following properties.

1. If $A \subseteq B$, then $0 \leq \mu(A) \leq \mu(B), \mu(\emptyset) = 0$.
2. $\mu(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$
3. $\mu([a, b]) = F(b) - F(a)$,
4. $\mu((a, b)) = F(b) - F(a)$
5. $\mu((a, b]) = F(b) - F(a)$
6. $\mu([a, b)) = F(b) - F(a)$.

**Proof:** First it is necessary to define the function $\mu$. This is contained in the following definition.

**Definition 19.8.4** For $A \subseteq \mathbb{R}$,
\[ \mu(A) = \inf \left\{ \sum_{j=1}^{\infty} (F(b_j) - F(a_j)) : A \subseteq \bigcup_{i=1}^{\infty} (a_i, b_i) \right\} \]
In words, you look at all coverings of $A$ with open intervals. For each of these open coverings, you add the “lengths” of the individual open intervals and you take the infimum of all such numbers obtained.

Then 1.) is obvious because if a countable collection of open intervals covers $B$, then it also covers $A$. Thus the set of numbers obtained for $B$ is smaller than the set of numbers for $A$. Why is $\mu(\emptyset) = 0$? Pick a point of continuity of $F$. Such points exist because $F$ is increasing and so it has
only countably many points of discontinuity. Let a be this point. Then \( \emptyset \subseteq (a - \delta, a + \delta) \) and so \( \mu(\emptyset) \leq F(a + \delta) - F(a - \delta) \) for every \( \delta > 0 \). Letting \( \delta \to 0 \), it follows that \( \mu(\emptyset) = 0 \).

Consider 2.). If any \( \mu(A_i) = \infty \), there is nothing to prove. The assertion simply is \( \infty \leq \infty \). Assume then that \( \mu(A_i) < \infty \) for all \( i \). Then for each \( m \in \mathbb{N} \) there exists a countable set of open intervals, \( \{(a^n_i, b^n_i)\}_{i=1}^\infty \) such that

\[
\mu(A_m) + \frac{\varepsilon}{2m} > \sum_{i=1}^\infty (F(b^n_i) - F(a^n_i)).
\]

Then using Theorem [\ref{thm:measure_zero}],

\[
\mu(\bigcup_{m=1}^\infty A_m) \leq \sum_{i,m} (F(b^n_i) - F(a^n_i)) = \sum_{m=1}^\infty \sum_{i=1}^\infty (F(b^n_i) - F(a^n_i))
\]

\[
\leq \sum_{m=1}^\infty \mu(A_m) + \frac{\varepsilon}{2m} = \sum_{m=1}^\infty \mu(A_m) + \varepsilon,
\]

and since \( \varepsilon \) is arbitrary, this establishes 2.).

Next consider 3.). By definition, there exists a sequence of open intervals, \( \{(a_i, b_i)\}_{i=1}^\infty \) whose union contains \([a, b]\) such that

\[
\mu([a, b]) + \varepsilon \geq \sum_{i=1}^\infty (F(b_i) - F(a_i))
\]

By Theorem [\ref{thm:countable}], finitely many of these intervals also cover \([a, b]\). It follows there exist finitely many of these intervals, denoted as \( \{(a_i, b_i)\}_{i=1}^n \), which overlap, such that \( a \in (a_1, b_1), b_1 \in (a_2, b_2), \ldots, b_n \in (a_n, b_n) \). Therefore,

\[
\mu([a, b]) \leq \sum_{i=1}^n (F(b_i) - F(a_i)).
\]

It follows

\[
\sum_{i=1}^n (F(b_i) - F(a_i)) \geq \mu([a, b])
\]

\[
\geq \sum_{i=1}^n (F(b_i) - F(a_i)) - \varepsilon
\]

\[
\geq F(b+) - F(a-) - \varepsilon
\]

Therefore, directly from the definition, since \([a, b] \subseteq (a - \delta, b + \delta)\)

\[
F(b + \delta) - F(a - \delta) \geq \mu([a, b]) \geq F(b+) - F(a-) - \varepsilon
\]

Letting \( \delta \to 0 \),

\[
F(b+) - F(a-) \geq \mu([a, b]) \geq F(b+) - F(a-) - \varepsilon
\]

Since \( \varepsilon \) is arbitrary, this shows

\[
\mu([a, b]) = F(b+) - F(a-),
\]

This establishes 3.).

Consider 4.). For small \( \delta > 0 \),

\[
\mu([a + \delta, b - \delta]) \leq \mu((a, b)) \leq \mu([a, b]).
\]

Therefore, from 3.) and the definition of \( \mu \),

\[
F((b - \delta)) - F((a + \delta)) \leq F((b - \delta) +) - F((a + \delta) -)
\]
\[ = \mu ([a + \delta, b - \delta]) \leq \mu (\{a, b\}) \leq F (b) - F (a+) \]

Now letting \( \delta \) decrease to 0 it follows

\[ F (b) - F (a+) \leq \mu (\{a, b\}) \leq F (b) - F (a+) \]

This shows 4.)

Consider 5.). From 3.) and 4.), for small \( \delta > 0, \)

\begin{align*}
F (b) - F (a+) & = \mu ([a + \delta, b]) \\
& \leq F (b) - F ((a + \delta) -) \\
& = \mu ([a + \delta, b]) \leq \mu ([a, b]) \\
& \leq F (b + \delta) + F (a+) \\
& \leq F (b) + F (a+) \\
\end{align*}

Now let \( \delta \) converge to 0 from above to obtain

\[ F (b) - F (a+) = \mu ([a, b]) = F (b) - F (a+) . \]

This establishes 5.) and 6.) is entirely similar to 5.). \( \square \)

The first two conditions of the above theorem are so important that we give something satisfying them a special name.

**Definition 19.8.5** Let \( \Omega \) be a nonempty set. A function mapping \( \mathcal{P} (\Omega) \rightarrow [0, \infty] \) is called an outer measure if it satisfies the following two condition.

1. If \( A \subseteq B \), then \( 0 \leq \mu (A) \leq \mu (B), \mu (\emptyset) = 0 \).
2. \( \mu (\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu (A_i) \)

### 19.9 Measures From Outer Measures

Earlier in Theorem 19.8.3 an outer measure on \( \mathcal{P} (\mathbb{R}) \) was constructed. This can be used to obtain a measure defined on \( \mathbb{R} \). However, the procedure for doing so is a special case of a general approach due to Caratheodory in about 1918.

**Definition 19.9.1** Let \( \Omega \) be a nonempty set and let \( \mu : \mathcal{P} (\Omega) \rightarrow [0, \infty] \) be an outer measure. For \( E \subseteq \Omega \), \( E \) is \( \mu \) measurable if for all \( S \subseteq \Omega \),

\[ \mu (S) = \mu (S \setminus E) + \mu (S \cap E) . \]  (19.11)

To help in remembering think of a measurable set \( E \), as a process which divides a given set into two pieces, the part in \( E \) and the part not in \( E \) as in \( \subseteq \). In the Bible, there are several incidents recorded in which a process of division resulted in more stuff than was originally present. \( \subseteq \) Measurable sets are exactly those which are incapable of such a miracle. You might think of the measurable sets as the non-miraculous sets. The idea is to show that they form a \( \sigma \) algebra on which the outer measure \( \mu \) is a measure.

First here is a definition and a lemma.

**Definition 19.9.2** \( (\mu | S) (A) = \mu (S \cap A) \) for all \( A \subseteq \Omega \). Thus \( \mu | S \) is the name of a new outer measure, called \( \mu \) restricted to \( S \).

---

1 Kings 17, 2 Kings 4, Mathew 14, and Mathew 15 all contain such descriptions. The stuff involved was either oil, bread, flour or fish. In mathematics such things have also been done with sets. In the book by Bruckner Bruckner and Thompson there is an interesting discussion of the Banach Tarski paradox which says it is possible to divide a ball in \( \mathbb{R}^3 \) into five disjoint pieces and assemble the pieces to form two disjoint balls of the same size as the first. The details can be found in: The Banach Tarski Paradox by Wagon, Cambridge University press. 1985. It is known that all such examples must involve the axiom of choice.
The next lemma indicates that the property of measurability is not lost by considering this restricted measure.

**Lemma 19.9.3** If \( A \) is \( \mu \) measurable, then \( A \) is \( \mu|_S \) measurable.

**Proof:** Suppose \( A \) is \( \mu \) measurable. It is desired to show that for all \( T \subseteq \Omega \),

\[
(\mu|_S)(T) = (\mu|_S)(T \cap A) + (\mu|_S)(T \setminus A).
\]

Thus it is desired to show

\[
\mu(S \cap T) = \mu(T \cap A \cap S) + \mu(T \cap S \cap A^C). \quad (19.12)
\]

But \( (19.12) \) holds because \( A \) is \( \mu \) measurable. Apply Definition \( (19.9) \) to \( S \cap T \) instead of \( S \).

If \( A \) is \( \mu|_S \) measurable, it does not follow that \( A \) is \( \mu \) measurable. Indeed, if you believe in the existence of non measurable sets, you could let \( A = S \) for such a \( \mu \) non measurable set and verify that \( S \) is \( \mu|_S \) measurable.

The next theorem is the main result on outer measures which shows that, starting with an outer measure, you can obtain a measure.

**Theorem 19.9.4** Let \( \Omega \) be a set and let \( \mu \) be an outer measure on \( \mathcal{P}(\Omega) \). The collection of \( \mu \) measurable sets \( S \), forms a \( \sigma \)-algebra and

\[
\text{If } F_i \in S, \ F_i \cap F_j = \emptyset, \text{ then } \mu(\bigcup_{i=1}^{\infty} F_i) = \sum_{i=1}^{\infty} \mu(F_i). \quad (19.13)
\]

If \( \cdots F_n \subseteq F_{n+1} \subseteq \cdots \), then if \( F = \bigcup_{n=1}^{\infty} F_n \) and \( F_n \in S \), it follows that

\[
\mu(F) = \lim_{n \to \infty} \mu(F_n). \quad (19.14)
\]

If \( \cdots F_n \supseteq F_{n+1} \supseteq \cdots \), and if \( F = \bigcap_{n=1}^{\infty} F_n \) for \( F_n \in S \) then if \( \mu(F_1) < \infty \),

\[
\mu(F) = \lim_{n \to \infty} \mu(F_n). \quad (19.15)
\]

This measure space is also complete which means that if \( \mu(F) = 0 \) for some \( F \in S \) then if \( G \subseteq F \), it follows \( G \in S \) also.

**Proof:** First note that \( \emptyset \) and \( \Omega \) are obviously in \( S \). Now suppose \( A, B \in S \). I will show \( A \setminus B \equiv A \cap B^C \) is in \( S \). To do so, consider the following picture.

It is required to show that

\[
\mu(S) = \mu(S \setminus (A \setminus B)) + \mu(S \cap (A \setminus B))
\]

First consider \( S \setminus (A \setminus B) \). From the picture, it equals

\[
(S \cap A^C \cap B^C) \cup (S \cap A \cap B) \cup (S \cap A^C \cap B)
\]

Therefore,

\[
\mu(S) \leq \mu(S \setminus (A \setminus B)) + \mu(S \cap (A \setminus B))
\]

\[
\leq \mu(S \cap A^C \cap B^C) + \mu(S \cap A \cap B) + \mu(S \cap A^C \cap B) + \mu(S \cap A \setminus B)
\]

\[
= \mu(S \cap A^C \cap B^C) + \mu(S \cap A \cap B) + \mu(S \cap A^C \cap B) + \mu(S \cap A \setminus B)
\]

\[
= \mu(S \cap A^C \cap B^C) + \mu(S \cap A \cap B) + \mu(S \cap A \cap B) + \mu(S \cap A^C \cap B)
\]

\[
= \mu(S \cap B^C) + \mu(S \cap B) = \mu(S)
\]
and so this shows that $A \setminus B \in S$ whenever $A, B \in S$.

Since $Ω \in S$, this shows that $A \in S$ if and only if $A^C \in S$. Now if $A, B \in S$, $A \cup B = (A^C \cap B^C)^C = (A^C \setminus B^C) \in S$. By induction, if $A_1, \ldots, A_n \in S$, then so is $\bigcup_{i=1}^n A_i$. If $A, B \in S$, with $A \cap B = ∅$, 

$$μ(A \cup B) = μ((A \cup B) \cap A) + μ((A \cup B) \setminus A) = μ(A) + μ(B).$$

By induction, if $A_i \cap A_j = ∅$ and $A_i \in S$,

$$μ(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n μ(A_i). \quad (19.16)$$

Now let $A = \bigcup_{i=1}^\infty A_i$ where $A_i \cap A_j = ∅$ for $i \neq j$.

$$\sum_{i=1}^\infty μ(A_i) ≥ μ(A) ≥ μ(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n μ(A_i).$$

Since this holds for all $n$, you can take the limit as $n \to ∞$ and conclude,

$$\sum_{i=1}^\infty μ(A_i) = μ(A)$$

which establishes 

Consider part (ii). Without loss of generality $μ(F_k) < ∞$ for all $k$ since otherwise there is nothing to show. Suppose $\{F_k\}$ is an increasing sequence of sets of $S$. Then letting $F_0 = ∅$, $\{F_{k+1} \setminus F_k\}_{k=0}^\infty$ is a sequence of disjoint sets of $S$ since it was shown above that the difference of two sets of $S$ is in $S$. Also note that from

$$μ(F_{k+1} \setminus F_k) + μ(F_k) = μ(F_{k+1})$$

and so if $μ(F_k) < ∞$, then

$$μ(F_{k+1} \setminus F_k) = μ(F_{k+1}) - μ(F_k).$$

Therefore, letting $F = \bigcup_{k=1}^\infty F_k$

which also equals

$$\bigcup_{k=1}^\infty (F_{k+1} \setminus F_k),$$

it follows from part (ii) just shown that

$$μ(F) = \sum_{k=0}^\infty μ(F_{k+1} \setminus F_k) = \lim_{n \to ∞} \sum_{k=0}^n μ(F_{k+1} \setminus F_k)$$

$$= \lim_{n \to ∞} \sum_{k=0}^n μ(F_{k+1}) - μ(F_k) = \lim_{n \to ∞} μ(F_{n+1}).$$

In order to establish (iii), let the $F_n$ be as given there. Then, since $(F_1 \setminus F_n)$ increases to $(F_1 \setminus F)$, (iii) implies

$$\lim_{n \to ∞} (μ(F_1) - μ(F_n)) = μ(F_1 \setminus F).$$

The problem is, I don’t know $F \in S$ and so it is not clear that $μ(F_1 \setminus F) = μ(F_1) - μ(F)$. However, $μ(F_1 \setminus F) + μ(F) ≥ μ(F_1)$ and so $μ(F_1 \setminus F) ≥ μ(F_1) - μ(F)$. Hence

$$\lim_{n \to ∞} (μ(F_1) - μ(F_n)) = μ(F_1 \setminus F) ≥ μ(F_1) - μ(F)$$

which implies

$$\lim_{n \to ∞} μ(F_n) ≤ μ(F).$$
But since $F \subseteq F_n$, 
\[ \mu(F) \leq \lim_{n \to \infty} \mu(F_n) \]
and this establishes \[\text{Infer}.\] Note that it was assumed $\mu(F_1) < \infty$ because $\mu(F_1)$ was subtracted from both sides.

It remains to show $S$ is closed under countable unions. Recall that if $A \in S$, then $A^c \in S$ and $S$ is closed under finite unions. Let $A_i \in S$, $A = \cup_{i=1}^n A_i$, $B_n = \cup_{i=1}^n A_i$. Then
\[ \mu(S) = \mu(S \cap B_n) + \mu(S \setminus B_n) \]
(19.17)
\[ = (\mu(S)(B_n) + (\mu(S)(B_n^C)). \]
By Lemma \[\text{Infer}.\] $B_n$ is $\mu(S)$ measurable and so is $B_n^C$. I want to show $\mu(S) \geq \mu(S \setminus A) + \mu(S \cap A)$. If $\mu(S) = \infty$, there is nothing to prove. Assume $\mu(S) < \infty$. Then apply Parts \[\text{Infer}.\] and \[\text{Infer}.\] to the outer measure $\mu(S)$ \[\text{Infer}.\] and let $n \to \infty$. Thus
\[ B_n \uparrow A, \quad B_n^C \downarrow A^c \]
and this yields $\mu(S) = (\mu(S)(A) + (\mu(S)(A^C) = \mu(S \cap A) + \mu(S \setminus A)$.

Therefore $A \in S$ and this proves Parts \[\text{Infer}.\] \[\text{Infer}.\] and \[\text{Infer}.\]

It only remains to verify the assertion about completeness. Letting $G$ and $F$ be as described above, let $S \subseteq \Omega$. I need to verify
\[ \mu(S) \geq \mu(S \cap G) + \mu(S \setminus G) \]
However,
\[ \mu(S \cap G) + \mu(S \setminus G) \leq \mu(S \cap F) + \mu(S \setminus F) + \mu(F \setminus G) \]
\[ = \mu(S \cap F) + \mu(S \setminus F) = \mu(S) \]
because by assumption, $\mu(F \setminus G) \leq \mu(F) = 0$. \[\square\]

**Corollary 19.9.5** Completeness is the same as saying that if $(E \setminus E') \cup (E' \setminus E) \subseteq N \in \mathcal{F}$ and $\mu(N) = 0$, then if $E \in \mathcal{F}$, it follows that $E' \in \mathcal{F}$ also.

**Proof:** If the new condition holds, then suppose $G \subseteq F$ where $\mu(F) = 0, F \in \mathcal{F}$. Then
\[ (G \setminus F) \cup (F \setminus G) \subseteq F \]
and $\mu(F)$ is given to equal 0. Therefore, $G \in \mathcal{F}$.

Now suppose the earlier version of completeness and let
\[ (E \setminus E') \cup (E' \setminus E) \subseteq N \in \mathcal{F} \]
where $\mu(N) = 0$ and $E \in \mathcal{F}$. Then we know
\[ (E \setminus E'), (E' \setminus E) \in \mathcal{F} \]
and all have measure zero. It follows $E \setminus (E \setminus E') = E \cap E' \in \mathcal{F}$. Hence
\[ E' = (E \cap E') \cup (E' \setminus E) \in \mathcal{F} \]
\[\square\]

The measure $\mu$ which results from the outer measure of Theorem \[Infer.\] is called the Lebesgue-Stieltjes measure associated with the integrator function $F$. Its properties will be discussed more in the next section.

Here is a general result. If you have a measure $\mu$, then by Proposition \[Infer.\] $\bar{\mu}$ defined there is an outer measure which agrees with $\mu$ on the $\sigma$ algebra of measurable sets $\mathcal{F}$. What of the measure determined by $\bar{\mu}$? Denoted still by $\bar{\mu}$. Is $\bar{\mu} = \mu$ on $\mathcal{F}$? Is $\mathcal{F}$ a subset of the $\bar{\mu}$ measurable sets, those which satisfy the criterion of being measurable? Suppose $E \in \mathcal{F}$. Is it the case that
\[ \bar{\mu}(S) = \bar{\mu}(S \setminus E) + \bar{\mu}(S \cap E)? \]
As usual, there is nothing to show if $\bar{\mu}(S) = \infty$ so assume this does not happen. Let $F \supseteq S, F \in \mathcal{F}$. Then by Proposition 19.6.2,
\[
\mu(F) = \mu(F \setminus E) + \mu(F \cap E)
\]
because everything is measurable. Then
\[
\bar{\mu}(S) = \inf_{F \supseteq S} \mu(F) \supseteq \inf_{F \supseteq S} (\mu(F \setminus E) + \mu(F \cap E))
\]
\[
\geq \inf_{F \supseteq S} \mu(F \setminus E) + \inf_{F \supseteq S} \mu(F \cap E)
\]
\[
\geq \inf_{F \supseteq S \setminus E} \mu(F \setminus E) + \inf_{F \supseteq S \cap E} \mu(F \cap E)
\]
\[
= \bar{\mu}(S \setminus E) + \bar{\mu}(S \cap E)
\]
Thus, indeed $\mathcal{F}$ is a subset of the $\bar{\mu}$ measurable sets. By Proposition 19.6.2, $\bar{\mu} = \mu$ on $\mathcal{F}$. This gives a way to complete a measure space which is described in the following proposition.

**Proposition 19.9.6** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $\bar{\mu}$ be the outer measure determined by $\mu$ as in Proposition 19.6.2. Also denote as $\mathcal{F}$, the $\sigma$ algebra of $\bar{\mu}$ measurable sets. Thus $(\Omega, \mathcal{F}, \bar{\mu})$ is a complete measure space in which $\mathcal{F} \supseteq \mathcal{F}$ and $\bar{\mu} = \mu$ on $\mathcal{F}$. Also, in this situation, if $\bar{\mu}(E) = 0$, then $E \in \mathcal{F}$. No new sets are obtained if $(\Omega, \mathcal{F}, \bar{\mu})$ is already complete.

**Proof:** If $S$ is a set,
\[
\bar{\mu}(S) \leq \bar{\mu}(S \cap E) + \bar{\mu}(S \setminus E)
\]
\[
\leq \bar{\mu}(E) + \bar{\mu}(S \setminus E)
\]
\[
= \bar{\mu}(S \setminus E) \leq \bar{\mu}(S)
\]
and so all inequalities are equal signs. Thus, if $E$ is a set with $\bar{\mu}(E) = 0$, then $E \in \mathcal{F}$. Suppose now that $(\Omega, \mathcal{F}, \mu)$ is complete. Let $F \in \mathcal{F}$. Then there exists $E \supseteq F$ such that $\mu(E) = \bar{\mu}(F)$. This is obvious if $\bar{\mu}(F) = \infty$. Otherwise, let $E_n \supseteq F, \mu(F) + \frac{1}{n} > \mu(E_n)$. Just let $E = \cap_n E_n$. Now $\bar{\mu}(E \setminus F) = 0$. Now also, there exists a set of $\mathcal{F}$ called $W$ such that $\mu(W) = 0$ and $W \supseteq E \setminus F$. Thus $E \setminus F \subseteq W$, a set of measure zero. Hence by completeness of $(\Omega, \mathcal{F}, \mu)$, it must be the case that $E \setminus F = E \cap F^C = G \in \mathcal{F}$. Then taking complements of both sides, $E^C \cap F = G^C \in \mathcal{F}$. Now take intersections with $E$. $F \in E \cap G^C \in \mathcal{F}$. $lacksquare$

### 19.10 One Dimensional Lebesgue Stieltjes Measure

Now with these major results about measures, it is time to specialize to the outer measure of Theorem 19.9.4. The next theorem gives Lebesgue Stieltjes measure on $\mathbb{R}$. The conditions 19.8.3 and 19.10.1 given below are known respectively as inner and outer regularity.

**Theorem 19.10.1** Let $\mathcal{F}$ denote the $\sigma$ algebra of Theorem 19.9.4, associated with the outer measure $\bar{\mu}$ in Theorem 19.8.3, on which $\mu$ is a measure. Then every open interval is in $\mathcal{F}$. So are all open and closed sets. Furthermore, if $E$ is any set in $\mathcal{F}$
\[
\mu(E) = \sup \{ \mu(K) : K \text{ compact, } K \subseteq E \} \tag{19.18}
\]
\[
\mu(E) = \inf \{ \mu(V) : V \text{ is an open set } V \supseteq E \} \tag{19.19}
\]

**Proof:** The first task is to show $(a, b) \in \mathcal{F}$. I need to show that for every $S \subseteq \mathbb{R}$,
\[
\mu(S) \geq \mu(S \cap (a, b)) + \mu\left(S \cap (a, b)^C\right) \tag{19.20}
\]
Suppose first \( S \) is an open interval, \((c, d)\). If \((c, d)\) has empty intersection with \((a, b)\) or is contained in \((a, b)\) there is nothing to prove. The above expression reduces to nothing more than \( \mu (S) = \mu (S) \). Suppose next that \((c, d) \supseteq (a, b)\). In this case, the right side of the above reduces to

\[
\begin{align*}
\mu ((a, b)) + \mu ((c, a] \cup [b, d)) & \leq F(b) - F(a+) + F(a+) - F(c+) + F(d) - F(b) \\
& = F(d) - F(c+) \equiv \mu ((c, d))
\end{align*}
\]

The only other cases are \( c \leq a < d \leq b \) or \( a \leq c < d \leq b \). Consider the first of these cases. Then the right side of (19.24) for \( S = (c, d) \) is

\[
\begin{align*}
\mu ((a, d)) + \mu ((c, a]) &= F(d) - F(a+) + F(a+) - F(c+) \\
& = F(d) - F(c+) = \mu ((c, d))
\end{align*}
\]

The last case is entirely similar. Thus (19.24) holds whenever \( S \) is an open interval. Now it is clear (19.24) also holds if \( \mu (S) = \infty \). Suppose then that \( \mu (S) < \infty \) and let

\[ S \subseteq \bigcup_{k=1}^{\infty} (a_k, b_k) \]

such that

\[ \mu (S) + \varepsilon > \sum_{k=1}^{\infty} (F(b_k) - F(a_k)) = \sum_{k=1}^{\infty} \mu ((a_k, b_k)). \]

Then since \( \mu \) is an outer measure, and using what was just shown,

\[
\begin{align*}
\mu (S \cap (a, b)) + \mu \left( S \cap (a, b)^c \right) & \leq \mu \left( \bigcup_{k=1}^{\infty} (a_k, b_k) \cap (a, b) \right) + \mu \left( \bigcup_{k=1}^{\infty} (a_k, b_k) \cap (a, b)^c \right) \\
& \leq \sum_{k=1}^{\infty} \mu ((a_k, b_k) \cap (a, b)) + \mu ((a_k, b_k) \cap (a, b)^c) \\
& \leq \sum_{k=1}^{\infty} \mu ((a_k, b_k)) \leq \mu (S) + \varepsilon.
\end{align*}
\]

Since \( \varepsilon \) is arbitrary, this shows (19.24) holds for any \( S \) and so any open interval is in \( \mathcal{F} \).

It follows any open set is in \( \mathcal{F} \). This follows from the fact that any open set in \( \mathbb{R} \) is the countable union of open intervals. See Theorem (19.23) for example. There can be no more than countably many disjoint open intervals because the rational numbers are dense and countable. Since each of these open intervals is in \( \mathcal{F} \) and \( \mathcal{F} \) is a \( \sigma \) algebra, their union is also in \( \mathcal{F} \). It follows every closed set is in \( \mathcal{F} \) also. This is because \( \mathcal{F} \) is a \( \sigma \) algebra and if a set is in \( \mathcal{F} \) then so is its complement. The closed sets are those which are complements of open sets. Then the regularity of this measure follows right away from Corollary (19.3.8) because the measure is finite on any open interval. \( \blacksquare \)

**Definition 19.10.2** When the integrator function is \( F(x) = x \), the Lebesgue Stieltjes measure just discussed is known as one dimensional Lebesgue measure and is denoted as \( m \).

**Proposition 19.10.3** For \( m \) Lebesgue measure, \( m ([a, b]) = m ((a, b)) = b - a \). Also \( m \) is translation invariant in the sense that if \( E \) is any Lebesgue measurable set, then \( m (x + E) = m (E) \).

**Proof:** The formula for the measure of an interval comes right away from Theorem (19.2.8). From this, it follows right away that whenever \( E \) is an interval, \( m (x + E) = m (E) \). Every open set is the countable disjoint union of open intervals by Theorem (19.2.8) so if \( E \) is an open set, then \( m (x + E) = m (E) \). What about closed sets? First suppose \( H \) is a closed and bounded set. Then letting \((-n, n) \supseteq H\),

\[
\mu (((-n, n) \setminus H) + x) + \mu (H + x) = \mu ((-n, n) + x)
\]
19.11. **Exercises**

1. Suppose you have \((X, \mathcal{F}, \mu)\) where \(\mathcal{F} \supseteq \mathcal{B}(X)\) and also \(\mu(B(x_0, r)) < \infty\) for all \(r > 0\). Let

   \[
   S(x_0, r) \equiv \{ x \in X : d(x, x_0) = r \}
   \]

   Show that

   \[
   \{ r > 0 : \mu(S(x_0, r)) > 0 \}
   \]

   cannot be uncountable. Explain why there exists a strictly increasing sequence \(r_n \to \infty\) such that \(\mu(\{ x : d(x, x_0) = r_n \}) = 0\). In other words, the skin of the ball has measure zero except for possibly countably many values of the radius \(r\).

2. In constructing Lebesgue Stieltjes measure on \(\mathbb{R}\), we defined the outer measure as

   \[
   \mu(A) \equiv \inf \left\{ \sum_{i=1}^{\infty} F(b_i-) - F(a_i+) : A \subseteq \bigcup_{i=1}^{\infty} (a_i, b_i) \right\}
   \]

   It was also shown that \(F(b-) - F(a+) = \mu((a, b))\). Show that this implies that \(\mu\) is outer regular on \(\mathcal{P}(\mathbb{R})\). That is, for any set

   \[
   A, \mu(A) = \inf \{ \mu(V) : V \supseteq A \text{ and } V \text{ is open} \}.
   \]

   In particular, this holds for all \(A \in \mathcal{F}\) the \(\sigma\) algebra of measurable sets. Now show that if \((X, \mu, \mathcal{F})\) is a measure space such that \(X\) is a complete separable metric space (Polish space) and if \(\mu\) is outer regular on \(\mathcal{F} \supseteq \mathcal{B}(X)\) and finite on every ball, then \(\mu\) must be inner regular on each set of \(\mathcal{F}\). That is,

   \[
   \mu(A) = \sup \{ \mu(K) : K \subseteq A \text{ and } K \text{ is compact} \}
   \]

   **Hint:** Let \(\{r_n\}\) be the increasing sequence of Problem \(\square\). Also let

   \[
   A_n = B(x_0, r_n), r_0 \equiv 0
   \]

   \[
   B_n = B(x_0, r_{n+1}). \text{ Thus } \overline{A_n} \subseteq B_n. \text{ Let } A \in \mathcal{F} \text{ and } A \subseteq A_n \subseteq \overline{A_n}. \text{ Then show that there exists an open set } V_n \supseteq \overline{A_n} \setminus A, V_n \subseteq B_n \text{ such that }
   \]

   \[
   \mu(V_n \setminus (\overline{A_n} \setminus A)) < \varepsilon
   \]
Then explain why $V_n^C \cap \overline{A_n} \subseteq A$ and $\mu (A \setminus (V_n^C \cap \overline{A_n})) < \varepsilon$. It might help to draw a picture on this last part. Thus there is a closed set $H$ contained in $A$ such that $\mu (A \setminus H) < \varepsilon$. Now recall the interesting result about regularity in Polish space. Thus there is $K$ compact such that $\mu (H \setminus K) < \varepsilon$. Of course $\mu$ is not finite but $\mu$ restricted to $B_n$ is. Now let $F$ be arbitrary. Then let $l < \mu (F)$ and argue that $l < \mu (F \cap B(x_0,r_n))$ for some $n$. Then use what was just shown.

3. Suppose you have any measure space $(\Omega, \mathcal{F}, \mu)$. The problem is that it might not be a complete measure space. That is, you might have $\mu (F) = 0$ and $G \subseteq F$ but $G \notin \mathcal{F}$. Define the following $\mu^*$ on $\mathcal{P} (\Omega)$.

$$\mu^* (F) = \inf \left\{ \sum_{i=1}^{\infty} \mu (E_i) : F \subseteq \bigcup_{i=1}^{\infty} E_i \right\}$$

Show first that $\mu^*$ is an outer measure. Next show that it agrees with $\mu$ on $\mathcal{F}$ and that for every $E \in \mathcal{F}$,

$$\mu^* (S) = \mu (S \cap F) + \mu^* (S \cap F^C)$$

From the Carathéodory procedure for constructing a measure space, there exists a $\sigma$ algebra $\tilde{\mathcal{F}}$ which contains $\mathcal{F}$ on which $\mu^*$ is a complete measure. This is called the completion of the measure space.

4. Consider the Cantor set. This is obtained by starting with $[0,1]$ deleting $1/3, 2/3, 3/3$ and then taking the two closed intervals which result and deleting the middle open third of each of these and continuing this way. Let $J_k$ denote the union of the $2^k$ closed intervals which result at the $k$th step of the construction. The Cantor set is $J \equiv \cap_{k=1}^{\infty} J_k$. Explain why $J$ is a nonempty compact subset of $\mathbb{R}$. Show that every point of $J$ is a limit point of $J$. Also show there exists a mapping from $J$ onto $[0,1]$ even though the sum of the lengths of the deleted open intervals is $1$. Show that the Cantor set has empty interior. If $x \in J$, consider the connected component of $x$. Show that this connected component is just $x$.

5. Lebesgue measure was discussed. Recall that $m ((a,b)) = b - a$ and it is defined on a $\sigma$ algebra which contains the Borel sets, more generally on $\mathcal{P} (\mathbb{R})$. Also recall that $m$ is translation invariant. Let $x \sim y$ if and only if $x - y \in \mathbb{Q}$. Show this is an equivalence relation. Now let $W$ be a set of positive measure which is contained in $(0,1)$. For $x \in W$, let $[x]$ denote those $y \in W$ such that $x \sim y$. Thus the equivalence classes partition $W$. Use axiom of choice to obtain a set $S \subseteq W$ such that $S$ consists of exactly one element from each equivalence class. Let $\mathbb{T}$ denote the rational numbers in $[-1,1]$. Consider $\mathbb{T} + S \subseteq [-1,2]$. Explain why $\mathbb{T} + S \supseteq W$. For $\mathbb{T} \equiv \{r_j\}$, explain why the sets $\{r_j + S\}_j$ are disjoint. Now suppose $S$ is measurable. Then show that you have a contradiction if $m (S) = 0$ since $m (W) > 0$ and you also have a contradiction if $m (S) > 0$ because $\mathbb{T} + S$ consists of countably many disjoint sets. Explain why $S$ cannot be measurable. Thus there exists $T \subseteq \mathbb{R}$ such that

$$m (T) < m (T \cap S) + m (T \cap S^C) .$$

Is there an open interval $(a,b)$ such that if $T = (a,b)$, then the above inequality holds?

6. Consider the following nested sequence of compact sets, $\{P_n\}$. Let $P_1 = [0,1]$, $P_2 = [0,1/3] \cup [2/3,1]$, etc. To go from $P_n$ to $P_{n+1}$, delete the open interval which is the middle third of each closed interval in $P_n$. Let $P = \cap_{n=1}^{\infty} P_n$. By the finite intersection property of compact sets, $P \neq \emptyset$. Show $m (P) = 0$. If you feel ambitious also show there is a one to one onto mapping of $[0,1]$ to $P$. The set $P$ is called the Cantor set. Thus, although $P$ has measure zero, it has the same number of points in it as $[0,1]$ in the sense that there is a one to one and onto mapping from one to the other. **Hint:** There are various ways of doing this last part but the most enlightenment is obtained by exploiting the topological properties of the Cantor set rather than some silly representation in terms of sums of powers of two and three. All you need to do is use the Schröder Bernstein theorem and show there is an onto map from the Cantor set.
11.1.38

Consider the sequence of functions defined in the following way. Let \( f_1(x) = x \) on \([0, 1]\). To get from \( f_n \) to \( f_{n+1} \), let \( f_{n+1} = f_n \) on all intervals where \( f_n \) is constant. If \( f_n \) is nonconstant on \([a, b]\), let \( f_{n+1}(a) = f_n(a) \), \( f_{n+1}(b) = f_n(b) \), \( f_{n+1} \) is piecewise linear and equal to \( \frac{1}{2}(f_n(a) + f_n(b)) \) on the middle third of \([a, b]\). Sketch a few of these and you will see the pattern. The process of modifying a nonconstant section of the graph of this function is illustrated in the following picture.

Show \( \{f_n\} \) converges uniformly on \([0, 1]\). If \( f(x) = \lim_{n \to \infty} f_n(x) \), show that \( f(0) = 0 \), \( f(1) = 1 \), \( f \) is continuous, and \( f'(x) = 0 \) for all \( x \notin P \) where \( P \) is the Cantor set of Problem 3. This function is called the Cantor function. It is a very important example to remember. Note it has derivative equal to zero a.e. and yet it succeeds in climbing from 0 to 1. Explain why this interesting function is not absolutely continuous although it is continuous. **Hint:** This isn’t too hard if you focus on getting a careful estimate on the difference between two successive functions in the list considering only a typical small interval in which the change takes place. The above picture should be helpful.

7. Consider the sequence of functions defined in the following way. Let \( f_1(x) = x \) on \([0, 1]\). To get from \( f_n \) to \( f_{n+1} \), let \( f_{n+1} = f_n \) on all intervals where \( f_n \) is constant. If \( f_n \) is nonconstant on \([a, b]\), let \( f_{n+1}(a) = f_n(a) \), \( f_{n+1}(b) = f_n(b) \), \( f_{n+1} \) is piecewise linear and equal to \( \frac{1}{2}(f_n(a) + f_n(b)) \) on the middle third of \([a, b]\). Sketch a few of these and you will see the pattern. The process of modifying a nonconstant section of the graph of this function is illustrated in the following picture.

Show \( \{f_n\} \) converges uniformly on \([0, 1]\). If \( f(x) = \lim_{n \to \infty} f_n(x) \), show that \( f(0) = 0 \), \( f(1) = 1 \), \( f \) is continuous, and \( f'(x) = 0 \) for all \( x \notin P \) where \( P \) is the Cantor set of Problem 3. This function is called the Cantor function. It is a very important example to remember. Note it has derivative equal to zero a.e. and yet it succeeds in climbing from 0 to 1. Explain why this interesting function is not absolutely continuous although it is continuous. **Hint:** This isn’t too hard if you focus on getting a careful estimate on the difference between two successive functions in the list considering only a typical small interval in which the change takes place. The above picture should be helpful.

8. **This problem gives a very interesting example found in the book by McShane [28].** Let \( g(x) = x + f(x) \) where \( f \) is the strange function of Problem 3. Let \( P \) be the Cantor set of Problem 3. Let \([0, 1] \setminus P = \bigcup_{j=1}^{\infty} I_j \) where \( I_j \) is open and \( I_j \cap I_k = \emptyset \) if \( j \neq k \). These intervals are the connected components of the complement of the Cantor set. Show \( m(g(I_j)) = m(I_j) \) so

\[
m(g(\bigcup_{j=1}^{\infty} I_j)) = \sum_{j=1}^{\infty} m(g(I_j)) = \sum_{j=1}^{\infty} m(I_j) = 1.
\]

Thus \( m(g(P)) = 1 \) because \( g([0, 1]) = [0, 2] \). By Problem 4 there exists a set, \( A \subseteq g(P) \) which is non measurable. Define \( \phi(x) = \chi_A(g(x)) \). Thus \( \phi(x) = 0 \) unless \( x \in P \). Tell why \( \phi \) is measurable. (Recall \( m(P) = 0 \) and Lebesgue measure is complete.) Now show that \( \chi_A(y) = \phi(g^{-1}(y)) \) for \( y \in [0, 2] \). Tell why \( g^{-1} \) is continuous but \( \phi \circ g^{-1} \) is not measurable. (This is an example of measurable \( \circ \) continuous \( \neq \) measurable.) Show there exist Lebesgue measurable sets which are not Borel measurable. **Hint:** The function, \( \phi \) is Lebesgue measurable. Now recall that Borel \( \circ \) measurable = measurable.

9. Show that every countable set of real numbers is of measure zero.

10. Review the Cantor set in Problem 3 on Page 245. You deleted middle third open intervals. Show that you can take out open intervals in the middle which are not necessarily middle thirds, and end up with a set \( C \) which has Lebesgue measure equal to \( 1 - \varepsilon \). Also show if you can that there exists a continuous and one to one map \( f : C \to J \) where \( J \) is the usual Cantor set of Problem 3 which also has measure 0.
Chapter 20

The Abstract Lebesgue Integral

The general Lebesgue integral requires a measure space, \((\Omega, \mathcal{F}, \mu)\) and, to begin with, a nonnegative measurable function. I will use Lemma 1.14.2 about interchanging two supremums frequently. Also, I will use the observation that if \(\{a_n\}\) is an increasing sequence of points of \([0, \infty]\), then \(\sup_n a_n = \lim_{n \to \infty} a_n\) which is obvious from the definition of \(\sup\). We have lots of good examples of measure spaces at this point, namely the Lebesgue-Stieltjes measures defined above. Included in this is one dimensional Lebesgue measure in which \(F(x) = x\). However, what follows is completely general, requiring only a measure space and measure.

20.1 Definition For Nonnegative Measurable Functions

20.1.1 Riemann Integrals For Decreasing Functions

First of all, the notation

\[
[g < f]
\]

is short for

\[
\{\omega \in \Omega : g(\omega) < f(\omega)\}
\]

with other variants of this notation being similar. Also, the convention, \(0 \cdot \infty = 0\) will be used to simplify the presentation whenever it is convenient to do so. The notation \(a \wedge b\) means the minimum of \(a\) and \(b\).

**Definition 20.1.1** Let \(f : [a, b] \to [0, \infty]\) be decreasing. Define

\[
\int_a^b f(\lambda) \, d\lambda \equiv \lim_{M \to \infty} \int_a^b M \wedge f(\lambda) \, d\lambda = \sup_M \int_a^b M \wedge f(\lambda) \, d\lambda
\]

where \(a \wedge b\) means the minimum of \(a\) and \(b\). Note that for \(f\) bounded,

\[
\sup_M \int_a^b M \wedge f(\lambda) \, d\lambda = \int_a^b f(\lambda) \, d\lambda
\]

where the integral on the right is the usual Riemann integral because eventually \(M > f\). For \(f\) a nonnegative decreasing function defined on \([0, \infty)\),

\[
\int_0^\infty f d\lambda \equiv \lim_{R \to \infty} \int_0^R f \, d\lambda = \sup_{R > 1} \int_0^R f \, d\lambda = \sup_{R > M > 0} \int_0^R f \wedge M \, d\lambda
\]

Since decreasing bounded functions are Riemann integrable, the above definition is well defined.

Now here are some obvious properties.
CHAPTER 20. THE ABSTRACT LEBESGUE INTEGRAL

Lemma 20.1.2 Let \( f \) be a decreasing nonnegative function defined on an interval \([a, b]\). Then if 
\[ [a, b] = \bigcup_{k=1}^{m} I_k \]
where \( I_k \equiv [a_k, b_k] \) and the intervals \( I_k \) are non overlapping, it follows 
\[
\int_{a}^{b} f \, d\lambda = \sum_{k=1}^{m} \int_{a_k}^{b_k} f \, d\lambda.
\]

Proof: This follows from the computation,
\[
\int_{a}^{b} f \, d\lambda = \lim_{M \to \infty} \int_{a}^{b} f \wedge M \, d\lambda = \lim_{M \to \infty} \sum_{k=1}^{m} \int_{a_k}^{b_k} f \wedge M \, d\lambda = \sum_{k=1}^{m} \int_{a_k}^{b_k} f \, d\lambda.
\]

Note both sides could equal \(+\infty\). ■

20.1.2 The Lebesgue Integral For Nonnegative Functions

Here is the definition of the Lebesgue integral of a function which is measurable and has values in \([0, \infty]\).

Definition 20.1.3 Let \((\Omega, \mathcal{F}, \mu)\) be a measure space and suppose \( f : \Omega \to [0, \infty] \) is measurable. Then define 
\[
\int f \, d\mu \equiv \int_{0}^{\infty} \mu ([f > \lambda]) \, d\lambda
\]
which makes sense because \( \lambda \to \mu ([f > \lambda]) \) is nonnegative and decreasing.

Note that if \( f \leq g \), then \( \int f \, d\mu \leq \int g \, d\mu \) because \( \mu ([f > \lambda]) \leq \mu ([g > \lambda]) \).

For convenience \( \sum_{i=1}^{\infty} a_i = 0 \).

Lemma 20.1.4 In the situation of the above definition,
\[
\int f \, d\mu = \sup_{h > 0} \sum_{i=1}^{\infty} \mu ([f > hi]) \, h
\]

Proof: Let \((h, R) \in \mathbb{N} \) satisfy \( R - h < hm(h, R) \leq R \). Then \( \lim_{R \to \infty} m(h, R) = \infty \) and so
\[
\int f \, d\mu = \int_{0}^{\infty} \mu ([f > \lambda]) \, d\lambda = \sup_{M} \sup_{R} \int_{0}^{hm(h, R)} \mu ([f > \lambda]) \wedge M \, d\lambda
\]
\[
= \sup_{M} \sup_{R} \left( \sum_{k=1}^{m(h, R)} \mu ([f > kh]) \wedge M \right) \, h
\]
The sum is just a lower sum for the integral \( \int_{0}^{hm(h, R)} \mu ([f > \lambda]) \wedge M \, d\lambda \). Hence, switching the order of the sups, this equals
\[
\sup_{R > 0} \sup_{h > 0} \sum_{k=1}^{m(h, R)} \mu ([f > kh]) \wedge M \right) \, h = \sup_{R > 0} \lim_{M \to \infty} \sum_{k=1}^{m(h, R)} \mu ([f > kh]) \wedge M \right) \, h
\]
\[
= \sup_{h > 0} \sum_{k=1}^{m(R, h)} \mu ([f > kh]) \, h = \sup_{h > 0} \sum_{k=1}^{\infty} \mu ([f > kh]) \, h \]
■
20.2 The Lebesgue Integral For Nonnegative Simple Functions

To begin with, here is a useful lemma.

**Lemma 20.2.1** If \( f(\lambda) = 0 \) for all \( \lambda > a \), where \( f \) is a decreasing nonnegative function, then
\[
\int_0^\infty f(\lambda) \, d\lambda = \int_0^a f(\lambda) \, d\lambda.
\]

**Proof:** From the definition,
\[
\int_0^\infty f(\lambda) \, d\lambda = \lim_{R \to \infty} \int_0^R f(\lambda) \, d\lambda = \sup_{R > 1} \int_0^R f(\lambda) \, d\lambda
\]
\[
= \sup_{R > 1} \sup_M \int_0^R f(\lambda) \, \wedge M \, d\lambda
\]
\[
= \sup_{M} \sup_{R > 1} \int_0^a f(\lambda) \, \wedge M \, d\lambda
\]
\[
= \sup_{M} \int_0^a f(\lambda) \, \wedge M \, d\lambda = \int_0^a f(\lambda) \, d\lambda.
\]

Now the Lebesgue integral for a nonnegative function has been defined, what does it do to a nonnegative simple function? Recall a nonnegative simple function is one which has finitely many nonnegative real values which it assumes on measurable sets. Thus a simple function can be written in the form
\[
s(\omega) = \sum_{i=1}^n c_i \chi_{E_i}(\omega)
\]
where the \( c_i \) are each nonnegative, the distinct values of \( s \).

**Lemma 20.2.2** Let \( s(\omega) = \sum_{i=1}^p a_i \chi_{E_i}(\omega) \) be a nonnegative simple function where the \( E_i \) are distinct but the \( a_i \) might not be. Then
\[
\int s \, d\mu = \sum_{i=1}^p a_i \mu(E_i).
\]  \hspace{1cm} (20.1)

**Proof:** Without loss of generality, assume \( 0 = a_0 < a_1 \leq a_2 \leq \cdots \leq a_p \) and that \( \mu(E_i) < \infty, i > 0 \). Here is why. If \( \mu(E_i) = \infty \), then letting \( a \in (a_{i-1}, a_i) \), by Lemma 20.2.1, the left side would be
\[
\int_0^\infty \mu([s > \lambda]) \, d\lambda \geq \int_{a_0}^{a_i} \mu([s > \lambda]) \, d\lambda
\]
\[
= \sup_M \int_0^{a_i} \mu([s > \lambda]) \, \wedge M \, d\lambda
\]
\[
= \sup_M M a_i = \infty
\]
and so both sides are equal to \( \infty \). Thus it can be assumed for each \( i, \mu(E_i) < \infty \). Then it follows from Lemma [41](#) and Lemma [42](#),
\[
\int_0^\infty \mu([s > \lambda]) \, d\lambda = \int_0^{a_p} \mu([s > \lambda]) \, d\lambda = \sum_{k=1}^p \int_{a_{k-1}}^{a_k} \mu([s > \lambda]) \, d\lambda
\]
\[
= \sum_{k=1}^p (a_k - a_{k-1}) \sum_{i=k}^p \mu(E_i) = \sum_{i=1}^p \mu(E_i) \sum_{k=1}^i (a_k - a_{k-1}) = \sum_{i=1}^p a_i \mu(E_i). \]
Lemma 20.2.3 If \(a, b \geq 0\) and if \(s\) and \(t\) are nonnegative simple functions, then
\[
\int (as + bt) \, d\mu = a \int sd\mu + b \int td\mu.
\]

Proof: Let
\[
s(\omega) = \sum_{i=1}^{n} \alpha_i \chi_{A_i}(\omega), \quad t(\omega) = \sum_{i=1}^{m} \beta_j \chi_{B_j}(\omega)
\]
where \(\alpha_i\) are the distinct values of \(s\) and the \(\beta_j\) are the distinct values of \(t\). Clearly \(as + bt\) is a nonnegative simple function because it has finitely many values on measurable sets. In fact,
\[
(as + bt)(\omega) = \sum_{j=1}^{m} \sum_{i=1}^{n} (a\alpha_i + b\beta_j) \chi_{A_i \cap B_j}(\omega)
\]
where the sets \(A_i \cap B_j\) are disjoint and measurable. By Lemma 20.2.2,
\[
\int (as + bt) \, d\mu = \sum_{j=1}^{m} \sum_{i=1}^{n} (a\alpha_i + b\beta_j) \mu(A_i \cap B_j)
\]
\[
= \sum_{i=1}^{n} a \sum_{j=1}^{m} \alpha_i \mu(A_i \cap B_j) + b \sum_{j=1}^{m} \beta_j \mu(A_i \cap B_j)
\]
\[
= a \sum_{i=1}^{n} \alpha_i \mu(A_i) + b \sum_{j=1}^{m} \beta_j \mu(B_j)
\]
\[
= a \int sd\mu + b \int td\mu. \quad \blacksquare
\]

### 20.3 The Monotone Convergence Theorem

The following is called the monotone convergence theorem. This theorem and related convergence theorems are the reason for using the Lebesgue integral.

**Theorem 20.3.1** (Monotone Convergence theorem) Let \(f\) have values in \([0, \infty]\) and suppose \(\{f_n\}\) is a sequence of nonnegative measurable functions having values in \([0, \infty]\) and satisfying
\[
\lim_{n \to \infty} f_n(\omega) = f(\omega) \text{ for each } \omega.
\]
\[
\cdots f_n(\omega) \leq f_{n+1}(\omega) \cdots
\]
Then \(f\) is measurable and
\[
\int f \, d\mu = \lim_{n \to \infty} \int f_n \, d\mu.
\]

Proof: By Lemma 20.2.3,
\[
\lim_{n \to \infty} \int f_n \, d\mu = \sup_n \int f_n \, d\mu
\]
\[
= \sup_n \sup_{h > 0} \sum_{k=1}^{\infty} \mu ([f_n > kh]) h = \sup_n \sup_{h > 0} \sum_{k=1}^{N} \mu ([f_n > kh]) h
\]
\[
= \sup_{h > 0} \sum_{k=1}^{N} \mu ([f > kh]) h = \sup_{h > 0} \sum_{k=1}^{\infty} \mu ([f > kh]) h = \int f \, d\mu. \quad \blacksquare
\]

To illustrate what goes wrong without the Lebesgue integral, consider the following example.
Example 20.3.2 Let \( \{r_n\} \) denote the rational numbers in [0, 1] and let

\[
f_n(t) = \begin{cases} 
1 & \text{if } t \notin \{r_1, \ldots, r_n\} \\
0 & \text{otherwise}
\end{cases}
\]

Then \( f_n(t) \uparrow f(t) \) where \( f \) is the function which is one on the rationals and zero on the irrationals. Each \( f_n \) is Riemann integrable (why?) but \( f \) is not Riemann integrable. Therefore, you can’t write \( \int f \, dx = \lim_{n\to\infty} \int f_n \, dx \). In fact, \( \int f_n \, dx = 0 \) for each \( n \).

A meta-mathematical observation related to this type of example is this. If you can choose your functions, you don’t need the Lebesgue integral. The Riemann Darboux integral is just fine. It is when you can’t choose your functions and they come to you as pointwise limits that you really need the superior Lebesgue integral or at least something more general than the Riemann integral. The Riemann integral is entirely adequate for evaluating the seemingly endless lists of boring problems found in calculus books. It is shown later that the two integrals coincide when the Lebesgue integral is taken with respect to Lebesgue measure and the function being integrated is Riemann integrable.

20.4 Other Definitions

To review and summarize the above, if \( f \geq 0 \) is measurable,

\[
\int f \, d\mu = \int_0^\infty \mu ([f > \lambda]) \, d\lambda \tag{20.2}
\]

another way to get the same thing for \( \int f \, d\mu \) is to take an increasing sequence of nonnegative simple functions, \( \{s_n\} \) with \( s_n(\omega) \to f(\omega) \) and then by monotone convergence theorem,

\[
\int f \, d\mu = \lim_{n \to \infty} \int s_n
\]

where if \( s_n(\omega) = \sum_{j=1}^m c_j \chi_{E_j}(\omega) \),

\[
\int s_n \, d\mu = \sum_{i=1}^m c_i \mu (E_i).
\]

Similarly this also shows that for such nonnegative measurable function,

\[
\int f \, d\mu = \sup \left\{ \int s : 0 \leq s \leq f, \ s \ \text{simple} \right\}
\]

Here is an equivalent definition of the integral of a nonnegative measurable function. The fact it is well defined has been discussed above.

Definition 20.4.1 For \( s \) a nonnegative simple function,

\[
s(\omega) = \sum_{k=1}^n c_k \chi_{E_k}(\omega), \quad \int s = \sum_{k=1}^n c_k \mu (E_k).
\]

For \( f \) a nonnegative measurable function,

\[
\int f \, d\mu = \sup \left\{ \int s : 0 \leq s \leq f, \ s \ \text{simple} \right\}.
\]
20.5 Fatou’s Lemma

The next theorem, known as Fatou’s lemma is another important theorem which justifies the use of the Lebesgue integral.

**Theorem 20.5.1 (Fatou’s lemma)** Let \( f_n \) be a nonnegative measurable function. Let \( g(\omega) = \lim \inf_{n \to \infty} f_n(\omega) \). Then \( g \) is measurable and

\[
\int g \, d\mu \leq \lim \inf_{n \to \infty} \int f_n \, d\mu.
\]

In other words,

\[
\int \left( \lim \inf_{n \to \infty} f_n \right) d\mu \leq \lim \inf_{n \to \infty} \int f_n \, d\mu
\]

**Proof:** Let \( g_n(\omega) = \inf \{ f_k(\omega) : k \geq n \} \). Then

\[
g_n^{-1}([a, \infty)) = \cap_{k=n}^{\infty} f_k^{-1}([a, \infty])
\]

\[
= (\cup_{k=n}^{\infty} f_k^{-1}([a, \infty]))^C \in \mathcal{F}.
\]

Thus \( g_n \) is measurable by Lemma 19.1.4. Also \( g(\omega) = \lim_{n \to \infty} g_n(\omega) \) so \( g \) is measurable because it is the pointwise limit of measurable functions. Now the functions \( g_n \) form an increasing sequence of nonnegative measurable functions so the monotone convergence theorem applies. This yields

\[
\int g \, d\mu = \lim_{n \to \infty} \int g_n \, d\mu \leq \lim \inf_{n \to \infty} \int f_n \, d\mu.
\]

The last inequality holding because

\[
\int g_n \, d\mu \leq \int f_n \, d\mu.
\]

(Note that it is not known whether \( \lim_{n \to \infty} \int f_n \, d\mu \) exists.) \( \blacksquare \)

20.6 The Integral’s Righteous Algebraic Desires

The monotone convergence theorem shows the integral wants to be linear. This is the essential content of the next theorem. We can’t say it is linear yet because to be linear, something must be defined on a vector space or something similar where it makes sense to consider linear combinations and the integral has only been defined at this point on nonnegative measurable functions.

**Theorem 20.6.1** Let \( f, g \) be nonnegative measurable functions and let \( a, b \) be nonnegative numbers. Then \( af + bg \) is measurable and

\[
\int (af + bg) \, d\mu = a \int f \, d\mu + b \int g \, d\mu.
\]  \hspace{1cm} (20.3)

**Proof:** By Theorem [19.1.6] on Page [464] there exist increasing sequences of nonnegative simple functions, \( s_n \to f \) and \( t_n \to g \). Then \( af + bg \), being the pointwise limit of the simple functions \( as_n + bt_n \), is measurable. Now by the monotone convergence theorem and Lemma 20.2.3,

\[
\int (af + bg) \, d\mu = \lim_{n \to \infty} \int as_n + bt_n \, d\mu
\]

\[
= \lim_{n \to \infty} \left( a \int s_n \, d\mu + b \int t_n \, d\mu \right)
\]

\[
= a \int f \, d\mu + b \int g \, d\mu. \blacksquare
\]

As long as you are allowing functions to take the value \(+\infty\), you cannot consider something like \( f + (-g) \) and so you can’t very well expect a satisfactory statement about the integral being linear until you restrict yourself to functions which have values in a vector space. To be linear, a function must be defined on a vector space. This is discussed next.
20.7 The Lebesgue Integral, $L^1$

The functions considered here have values in $\mathbb{C}$, which is a vector space. A function $f$ with values in $\mathbb{C}$ is of the form $f = \text{Re} f + i \text{Im} f$ where $\text{Re} f$ and $\text{Im} f$ are real valued functions. In fact

$$\text{Re} f = \frac{f + \overline{f}}{2}, \quad \text{Im} f = \frac{f - \overline{f}}{2i}.$$ 

Definition 20.7.1 Let $(Ω, S, μ)$ be a measure space and suppose $f : Ω → \mathbb{C}$. Then $f$ is said to be measurable if both $\text{Re} f$ and $\text{Im} f$ are measurable real valued functions.

Of course there is another definition of measurability which says that inverse images of measurable sets are measurable. This is equivalent to this new definition.

Lemma 20.7.2 Let $f : Ω → \mathbb{C}$. Then $f$ is measurable if and only if $\text{Re} f, \text{Im} f$ are both real valued measurable functions. Also if $f, g$ are complex measurable functions and $a, b$ are complex scalars, then $af + bg$ is also measurable.

Proof: \(⇒\) Suppose first that $f$ is measurable. Recall that $\mathbb{C}$ is considered as $\mathbb{R}^2$ with $(x, y)$ being identified with $x + iy$. Thus the open sets of $\mathbb{C}$ can be obtained with either of the two equivalent norms $|z| \equiv \sqrt{(\text{Re} z)^2 + (\text{Im} z)^2}$ or $|z|_\infty = \max (\text{Re} z, \text{Im} z)$. Therefore, if $f$ is measurable

$$\text{Re} f^{-1} (a, b) \cap \text{Im} f^{-1} (c, d) = f^{-1} ((a, b) + i (c, d)) \in \mathcal{F}.$$ 

In particular, you could let $(c, d) = \mathbb{R}$ and conclude that $\text{Re} f$ is measurable because in this case, the above reduces to the statement that $\text{Re} f^{-1} (a, b) \in \mathcal{F}$. Similarly $\text{Im} f$ is measurable.

\(⇐\) Next, if each of $\text{Re} f$ and $\text{Im} f$ are measurable, then

$$f^{-1} ((a, b) + i (c, d)) = \text{Re} f^{-1} (a, b) \cap \text{Im} f^{-1} (c, d) \in \mathcal{F}$$

and so, since every open set is the countable union of sets of the form $(a, b) + i (c, d)$, it follows that $f$ is measurable.

Now consider the last claim. Let

$$h : \mathbb{C} \times \mathbb{C} → \mathbb{C}$$

be given by $h (z, w) \equiv az + bw$. Then $h$ is continuous. If $f, g$ are complex valued measurable functions, consider the complex valued function,

$$h \circ (f, g) : Ω → \mathbb{C}$$

Then

$$(h \circ (f, g))^{-1} (\text{open}) = (f, g)^{-1} (h^{-1} (\text{open})) = (f, g)^{-1} (\text{open})$$

Now letting $U, V$ be open in $\mathbb{C},$

$$(f, g)^{-1} (U \times V) = f^{-1} (U) \cap g^{-1} (V) \in \mathcal{F}.$$ 

Since every open set in $\mathbb{C} \times \mathbb{C}$ is the countable union of sets of the form $U \times V$, it follows that $(f, g)^{-1} (\text{open})$ is in $\mathcal{F}$. Thus $af + bg$ is also complex measurable. \(\blacksquare\)

As is always the case for complex numbers, $|z|^2 = (\text{Re} z)^2 + (\text{Im} z)^2$. Also, for $g$ a real valued function, one can consider its positive and negative parts defined respectively as

$$g^+ (x) = \frac{g(x) + |g(x)|}{2}, \quad g^- (x) = \frac{|g(x)| - g(x)}{2}.$$ 

Thus $|g| = g^+ + g^-$ and $g = g^+ - g^-$ and both $g^+$ and $g^-$ are measurable nonnegative functions if $g$ is measurable.

Then the following is the definition of what it means for a complex valued function $f$ to be in $L^1 (Ω)$. 

**Definition 20.7.3** Let $(Ω, F, μ)$ be a measure space. Then a complex valued measurable function $f$ is in $L^1(Ω)$ if

$$\int |f| \, dμ < \infty.$$  

For a function in $L^1(Ω)$, the integral is defined as follows.

$$\int f \, dμ \equiv \int (\text{Re } f)^+ \, dμ - \int (\text{Re } f)^- \, dμ + i \left[ \int (\text{Im } f)^+ \, dμ - \int (\text{Im } f)^- \, dμ \right]$$

I will show that with this definition, the integral is linear and well defined. First note that it is clearly well defined because all the above integrals are of nonnegative functions and are each equal to a nonnegative real number because for $h$ equal to any of the functions, $|h| \leq |f|$ and $\int |f| \, dμ < \infty$.

Here is a lemma which will make it possible to show the integral is linear.

**Lemma 20.7.4** Let $g, h, g', h'$ be nonnegative measurable functions in $L^1(Ω)$ and suppose that $g - h = g' - h'$.

Then

$$\int g \, dμ - \int h \, dμ = \int g' \, dμ - \int h' \, dμ.$$  

**Proof:** By assumption, $g + h' = g' + h$. Then from the Lebesgue integral’s righteous algebraic desires, Theorem 20.6.1,  

$$\int g \, dμ + \int h' \, dμ = \int g' \, dμ + \int h' \, dμ$$

which implies the claimed result. ■

**Lemma 20.7.5** Let $\text{Re } (L^1(Ω))$ denote the vector space of real valued functions in $L^1(Ω)$ where the field of scalars is the real numbers. Then $\int dμ$ is linear on $\text{Re } (L^1(Ω))$, the scalars being real numbers.

**Proof:** First observe that from the definition of the positive and negative parts of a function,

$$(f + g)^+ - (f + g)^- = f^+ + g^+ - (f^- + g^-)$$

because both sides equal $f + g$. Therefore from Lemma 20.7.4 and the definition, it follows from Theorem 20.6.1 that

$$\int f + g \, dμ = \int (f + g)^+ - (f + g)^- \, dμ = \int f^+ + g^+ \, dμ - \int f^- + g^- \, dμ$$

$$= \int f^+ \, dμ + \int g^+ \, dμ - \left( \int f^- \, dμ + \int g^- \, dμ \right) = \int f \, dμ + \int g \, dμ.$$  

what about taking out scalars? First note that if $a$ is real and nonnegative, then $(af)^+ = af^+$ and $(af)^- = af^-$ while if $a < 0$, then $(af)^+ = -af^-$ and $(af)^- = -af^+$. These claims follow immediately from the above definitions of positive and negative parts of a function. Thus if $a < 0$ and $f \in L^1(Ω)$, it follows from Theorem 20.6.1 that

$$\int af \, dμ = \int (af)^+ \, dμ - \int (af)^- \, dμ = \int (-a) f^- \, dμ - \int (-a) f^+ \, dμ$$

$$= -a \int f^- \, dμ + a \int f^+ \, dμ = a \left( \int f^+ \, dμ - \int f^- \, dμ \right) = a \int f \, dμ.$$  

The case where $a \geq 0$ works out similarly but easier. ■

Now here is the main result.
Theorem 20.7.6 \( \int d\mu \) is linear on \( L^1(\Omega) \) and \( L^1(\Omega) \) is a complex vector space. If \( f \in L^1(\Omega) \), then \( \text{Re} f, \text{Im} f, \) and \( |f| \) are all in \( L^1(\Omega) \). Furthermore, for \( f \in L^1(\Omega) \),

\[
\int f d\mu = \int (\text{Re} f)^+ d\mu - \int (\text{Re} f)^- d\mu + i \left[ \int (\text{Im} f)^+ d\mu - \int (\text{Im} f)^- d\mu \right]
\]

and the triangle inequality holds,

\[
\left| \int f d\mu \right| \leq \int |f| d\mu. \tag{20.4}
\]

Also, for every \( f \in L^1(\Omega) \) it follows that for every \( \varepsilon > 0 \) there exists a simple function \( s \) such that \( |s| \leq |f| \) and

\[
\int |f - s| d\mu < \varepsilon.
\]

Also \( L^1(\Omega) \) is a vector space.

**Proof:** First consider the claim that the integral is linear. It was shown above that the integral is linear on \( \text{Re} \left( L^1(\Omega) \right) \). Then letting \( a + ib, c + id \) be scalars and \( f, g \) functions in \( L^1(\Omega) \),

\[
(a + ib) f + (c + id) g = (a + ib) (\text{Re} f + i \text{Im} f) + (c + id) (\text{Re} g + i \text{Im} g)
\]

\[= c \text{Re} (g) - b \text{Im} (f) - d \text{Im} (g) + a \text{Re} (f) + i (b \text{Re} (f) + c \text{Im} (g) + a \text{Im} (f) + d \text{Re} (g))\]

It follows from the definition that

\[
\int (a + ib) f + (c + id) g d\mu = \int (c \text{Re} (g) - b \text{Im} (f) - d \text{Im} (g) + a \text{Re} (f)) d\mu
\]

\[+ i \int (b \text{Re} (f) + c \text{Im} (g) + a \text{Im} (f) + d \text{Re} (g)) \] \[\tag{20.5}\]

Also, from the definition,

\[
(a + ib) \int f d\mu + (c + id) \int g d\mu = (a + ib) \left( \int \text{Re} f d\mu + i \int \text{Im} f d\mu \right)
\]

\[+ (c + id) \left( \int \text{Re} g d\mu + i \int \text{Im} g d\mu \right)\]

which equals

\[
= a \int \text{Re} f d\mu - b \int \text{Im} f d\mu + ib \int \text{Re} f d\mu + ia \int \text{Im} f d\mu
\]

\[+ c \int \text{Re} g d\mu - d \int \text{Im} g d\mu + id \int \text{Re} g d\mu - d \int \text{Im} g d\mu.\]

Using Lemma 41 and collecting terms, it follows that this reduces to 41. Thus the integral is linear as claimed.

Consider the claim about approximation with a simple function. Letting \( h \) equal any of \( (\text{Re} f)^+, (\text{Re} f)^-, (\text{Im} f)^+, (\text{Im} f)^- \),

\[
(20.6)
\]

It follows from the monotone convergence theorem and Theorem on Page 464 there exists a nonnegative simple function \( s \leq h \) such that

\[
\int |h - s| d\mu < \frac{\varepsilon}{4}.
\]
Therefore, letting $s_1, s_2, s_3, s_4$ be such simple functions, approximating respectively the functions listed in \ref{20.6}, and $s \equiv s_1 - s_2 + i (s_3 - s_4)$,

$$
\int |f - s| \, d\mu \leq \int \left( |(\text{Re } f)^+ - s| + \int |(\text{Re } f)^- - s| \right) \, d\mu \\
+ \int \left( |(\text{Im } f)^+ - s_3| + \int |(\text{Im } f)^- - s_4| \right) \, d\mu < \varepsilon
$$

It is clear from the construction that $|s| \leq |f|$.

What about \ref{20.4}? Let $\theta \in \mathbb{C}$ be such that $|\theta| = 1$ and $\theta \int f \, d\mu = |\int f \, d\mu|$. Then from what was shown above about the integral being linear,

$$
\int f \, d\mu = \theta \int f \, d\mu = \int \theta f \, d\mu = \int \text{Re} (\theta f) \, d\mu \leq \int |f| \, d\mu.
$$

If $f, g \in L^1(\Omega)$, then it is known that for $a, b$ scalars, it follows that $af + bg$ is measurable. See Lemma \ref{20.7.2}. Also

$$
\int |af + bg| \, d\mu \leq \int |a| |f| + |b| |g| \, d\mu < \infty. \quad \blacksquare
$$

The following corollary follows from this. The conditions of this corollary are sometimes taken as a definition of what it means for a function $f$ to be in $L^1(\Omega)$.

**Corollary 20.7.7** $f \in L^1(\Omega)$ if and only if there exists a sequence of complex simple functions, \{sn\} such that

$$
s_n(\omega) \to f(\omega) \quad \text{for all } \omega \in \Omega \quad \text{(20.7)}
$$

When $f \in L^1(\Omega)$,

$$
\int f \, d\mu = \lim_{n \to \infty} \int s_n. \quad \text{(20.8)}
$$

**Proof:** From the above theorem, if $f \in L^1$ there exists a sequence of simple functions \{sn\} such that

$$
\int |f - s_n| \, d\mu < 1/n, \quad s_n(\omega) \to f(\omega) \quad \text{for all } \omega
$$

Then

$$
\int |s_n - s_m| \, d\mu \leq \int |s_n - f| \, d\mu + \int |f - s_m| \, d\mu \leq \frac{1}{n} + \frac{1}{m}.
$$

Next suppose the existence of the approximating sequence of simple functions. Then $f$ is measurable because its real and imaginary parts are the limit of measurable functions. By Fatou’s lemma,

$$
\int |f| \, d\mu \leq \lim \inf_{n \to \infty} \int |s_n| \, d\mu < \infty
$$

because

$$
\left| \int |s_n| \, d\mu - \int |s_m| \, d\mu \right| \leq \int |s_n - s_m| \, d\mu
$$

which is given to converge to 0. Hence $\{ \int |s_n| \, d\mu \}$ is a Cauchy sequence and is therefore, bounded.

In case $f \in L^1(\Omega)$, letting \{sn\} be the approximating sequence, Fatou’s lemma implies

$$
\left| \int f \, d\mu - \int s_n \, d\mu \right| \leq \int |f - s_n| \, d\mu \leq \lim \inf_{m \to \infty} \int |s_m - s_n| \, d\mu < \varepsilon
$$

provided $n$ is large enough. Hence \ref{20.8} follows. \quad \blacksquare

This is a good time to observe the following fundamental observation which follows from a repeat of the above arguments.
Theorem 20.7.8 Suppose \( \Lambda (f) \in [0, \infty] \) for all nonnegative measurable functions and suppose that for \( a, b \geq 0 \) and \( f, g \) nonnegative measurable functions,

\[
\Lambda (af + bg) = a \Lambda (f) + b \Lambda (g).
\]

In other words, \( \Lambda \) wants to be linear. Then \( \Lambda \) has a unique linear extension to the set of measurable functions

\[
\{ f \text{ measurable} : \Lambda (|f|) < \infty \},
\]

this set being a vector space.

20.8 The Dominated Convergence Theorem

One of the major theorems in this theory is the dominated convergence theorem. Before presenting it, here is a technical lemma about \( \lim \sup \) and \( \lim \inf \) which is really pretty obvious from the definition.

Lemma 20.8.1 Let \( \{a_n\} \) be a sequence in \([-\infty, \infty] \). Then \( \lim_{n \to \infty} a_n \) exists if and only if

\[
\liminf_{n \to \infty} a_n = \limsup_{n \to \infty} a_n
\]

and in this case, the limit equals the common value of these two numbers.

Proof: Suppose first \( \lim_{n \to \infty} a_n = a \in \mathbb{R} \). Then, letting \( \varepsilon > 0 \) be given, \( a_n \in (a - \varepsilon, a + \varepsilon) \) for all \( n \) large enough, say \( n \geq N \). Therefore, both inf \( \{a_k : k \geq n\} \) and sup \( \{a_k : k \geq n\} \) are contained in \([a - \varepsilon, a + \varepsilon]\) whenever \( n \geq N \). It follows \( \limsup_{n \to \infty} a_n \) and \( \liminf_{n \to \infty} a_n \) are both in \([a - \varepsilon, a + \varepsilon]\), showing

\[
\left| \liminf_{n \to \infty} a_n - \limsup_{n \to \infty} a_n \right| < 2\varepsilon.
\]

Since \( \varepsilon \) is arbitrary, the two must be equal and they both must equal \( a \). Next suppose \( \lim_{n \to \infty} a_n = \infty \). Then if \( l \in \mathbb{R} \), there exists \( N \) such that for \( n \geq N \),

\[
l \leq a_n
\]

and therefore, for such \( n \),

\[
l \leq \inf \{a_k : k \geq n\} \leq \sup \{a_k : k \geq n\}
\]

and this shows, since \( l \) is arbitrary that

\[
\liminf_{n \to \infty} a_n = \limsup_{n \to \infty} a_n = \infty.
\]

The case for \(-\infty\) is similar.

Conversely, suppose \( \liminf_{n \to \infty} a_n = \limsup_{n \to \infty} a_n = a \). Suppose first that \( a \in \mathbb{R} \). Then, letting \( \varepsilon > 0 \) be given, there exists \( N \) such that if \( n \geq N \),

\[
\sup \{a_k : k \geq n\} - \inf \{a_k : k \geq n\} < \varepsilon
\]

therefore, if \( k, m > N \), and \( a_k > a_m \),

\[
|a_k - a_m| = a_k - a_m \leq \sup \{a_k : k \geq n\} - \inf \{a_k : k \geq n\} < \varepsilon
\]

showing that \( \{a_n\} \) is a Cauchy sequence. Therefore, it converges to \( a \in \mathbb{R} \), and as in the first part, the \( \lim \inf \) and \( \lim \sup \) both equal \( a \). If \( \liminf_{n \to \infty} a_n = \limsup_{n \to \infty} a_n = \infty \), then given \( l \in \mathbb{R} \), there exists \( N \) such that for \( n \geq N \),

\[
\inf_{n \geq N} a_n > l.
\]

Therefore, \( \lim_{n \to \infty} a_n = \infty \). The case for \(-\infty\) is similar.

Here is the dominated convergence theorem.
Theorem 20.8.2 (Dominated Convergence theorem) Let \( f_n \in L^1(\Omega) \) and suppose

\[
f(\omega) = \lim_{n \to \infty} f_n(\omega),
\]

and there exists a measurable function \( g \), with values in \([0, \infty]\), such that

\[
|f_n(\omega)| \leq g(\omega) \quad \text{and} \quad \int g(\omega) d\mu < \infty.
\]

Then \( f \in L^1(\Omega) \) and

\[
0 = \lim_{n \to \infty} \int |f_n - f| d\mu = \lim_{n \to \infty} \left| \int f d\mu - \int f_n d\mu \right|
\]

Proof: \( f \) is measurable by Theorem 19.1.2. Since \( |f| \leq g \), it follows that

\[
f \in L^1(\Omega) \quad \text{and} \quad |f - f_n| \leq 2g.
\]

By Fatou’s lemma (Theorem 20.5.1),

\[
\int 2g d\mu \leq \lim \inf_{n \to \infty} \int 2g - |f - f_n| d\mu

= \int 2g d\mu - \lim \inf_{n \to \infty} \int |f - f_n| d\mu.
\]

 Subtracting \( \int 2g d\mu \),

\[
0 \leq - \lim \sup_{n \to \infty} \int |f - f_n| d\mu.
\]

Hence

\[
0 \geq \lim \sup_{n \to \infty} \left( \int |f - f_n| d\mu \right)

\geq \lim \inf_{n \to \infty} \left( \int |f - f_n| d\mu \right) \geq \left| \int f d\mu - \int f_n d\mu \right| \geq 0.
\]

This proves the theorem by Lemma 20.8.1 because the \( \lim \sup \) and \( \lim \inf \) are equal. \( \blacksquare \)

Corollary 20.8.3 Suppose \( f_n \in L^1(\Omega) \) and \( f(\omega) = \lim_{n \to \infty} f_n(\omega) \). Suppose also there exist measurable functions, \( g_n, g \) with values in \([0, \infty]\) such that \( \lim_{n \to \infty} \int g_n d\mu = \int g d\mu, g_n(\omega) \to g(\omega) \) \( \mu \) a.e. and both \( \int g_n d\mu \) and \( \int g d\mu \) are finite. Also suppose \( |f_n(\omega)| \leq g_n(\omega) \). Then

\[
\lim_{n \to \infty} \int |f - f_n| d\mu = 0.
\]

Proof: It is just like the above. This time \( g + g_n - |f - f_n| \geq 0 \) and so by Fatou’s lemma,

\[
\int 2g d\mu - \lim \sup_{n \to \infty} \int |f - f_n| d\mu =

\lim \inf_{n \to \infty} \int (g_n + g) d\mu - \lim \sup_{n \to \infty} \int |f - f_n| d\mu

= \lim \inf_{n \to \infty} \int ((g_n + g) - |f - f_n|) d\mu \geq \int 2g d\mu
\]

and so \( - \lim \sup_{n \to \infty} \int |f - f_n| d\mu \geq 0 \). Thus

\[
0 \geq \lim \sup_{n \to \infty} \left( \int |f - f_n| d\mu \right)

\geq \lim \inf_{n \to \infty} \left( \int |f - f_n| d\mu \right) \geq \left| \int f d\mu - \int f_n d\mu \right| \geq 0.
\]

\footnote{Note that, since \( g \) is allowed to have the value \( \infty \), it is not known that \( g \in L^1(\Omega) \).}
Define 20.8.4 Let $E$ be a measurable subset of $\Omega$.

$$\int_E f d\mu \equiv \int f\chi_E d\mu.$$  

If $L^1(E)$ is written, the $\sigma$ algebra is defined as

$$\{E \cap A : A \in \mathcal{F}\}$$

and the measure is $\mu$ restricted to this smaller $\sigma$ algebra. Clearly, if $f \in L^1(\Omega)$, then

$$f\chi_E \in L^1(E)$$

and if $f \in L^1(E)$, then letting $\tilde{f}$ be the 0 extension of $f$ off of $E$, it follows $\tilde{f} \in L^1(\Omega)$.

20.9 Exercises

1. Let $\Omega = \mathbb{N} = \{1, 2, \ldots\}$. Let $\mathcal{F} = \mathcal{P}(\mathbb{N})$, the set of all subsets of $\mathbb{N}$, and let $\mu(S) =$ number of elements in $S$. Thus $\mu(\{1\}) = 1 = \mu(\{2\})$, $\mu(\{1, 2\}) = 2$, etc. In this case, all functions are measurable. For a nonnegative function, $f$ defined on $\mathbb{N}$, show

$$\int f d\mu = \sum_{k=1}^{\infty} f(k)$$

What do the monotone convergence and dominated convergence theorems say about this example?

2. For the measure space of Problem 1, give an example of a sequence of nonnegative measurable functions $f_n$ converging pointwise to a function $f$, such that inequality is obtained in Fatou’s lemma.

3. If $(\Omega, \mathcal{F}, \mu)$ is a measure space and $f \geq 0$ is measurable, show that if $g(\omega) = f(\omega)$ a.e. $\omega$ and $g \geq 0$, then $\int g d\mu = \int f d\mu$. Show that if $f, g \in L^1(\Omega)$ and $g(\omega) = f(\omega)$ a.e. then $\int g d\mu = \int f d\mu$.

4. Let $\{f_n\}, f$ be measurable functions with values in $\mathbb{C}$. $\{f_n\}$ converges in measure if

$$\lim_{n \to \infty} \mu(\{x \in \Omega : |f(x) - f_n(x)| \geq \varepsilon\}) = 0$$

for each fixed $\varepsilon > 0$. Prove the theorem of F. Riesz. If $f_n$ converges to $f$ in measure, then there exists a subsequence $\{f_{n_k}\}$ which converges to $f$ a.e. In case $\mu$ is a probability measure, this is called convergence in probability. It does not imply pointwise convergence but does imply that there is a subsequence which converges pointwise off a set of measure zero. Hint: Choose $n_1$ such that

$$\mu(x : |f(x) - f_{n_1}(x)| \geq 1) < 1/2.$$  

Choose $n_2 > n_1$ such that

$$\mu(x : |f(x) - f_{n_2}(x)| \geq 1/2) < 1/2^2,$$

$n_3 > n_2$ such that

$$\mu(x : |f(x) - f_{n_3}(x)| \geq 1/3) < 1/2^3,$$

e tc. Now consider what it means for $f_{n_k}(x)$ to fail to converge to $f(x)$. Use the Borel Cantelli lemma of Problem 14 on Page 472.
5. Suppose \((\Omega, \mu)\) is a finite measure space \((\mu(\Omega) < \infty)\) and \(\mathcal{S} \subseteq L^1(\Omega)\). Then \(\mathcal{S}\) is said to be uniformly integrable if for every \(\varepsilon > 0\) there exists \(\delta > 0\) such that if \(E\) is a measurable set satisfying \(\mu(E) < \delta\), then

\[
\int_E |f| \, d\mu < \varepsilon
\]

for all \(f \in \mathcal{S}\). Show \(\mathcal{S}\) is uniformly integrable and bounded in \(L^1(\Omega)\) if there exists an increasing function \(h\) which satisfies

\[
\lim_{t \to \infty} \frac{h(t)}{t} = \infty, \quad \sup \left\{ \int_\Omega h(|f|) \, d\mu : f \in \mathcal{S} \right\} < \infty.
\]

\(\mathcal{S}\) is bounded if there is some number, \(M\) such that

\[
\int |f| \, d\mu \leq M
\]

for all \(f \in \mathcal{S}\).

6. A collection \(\mathcal{S} \subseteq L^1(\Omega), (\Omega, \mathcal{F}, \mu)\) a finite measure space, is called equiintegrable if for every \(\varepsilon > 0\) there exists \(\lambda > 0\) such that

\[
\int_{\{|f| \geq \lambda\}} |f| \, d\mu < \varepsilon
\]

for all \(f \in \mathcal{S}\). Show that \(\mathcal{S}\) is equiintegrable, if and only if it is uniformly integrable and bounded. The equiintegrable condition is pretty popular in probability.

7. There is a general construction called product measure. You have two finite measure spaces.

\((X, \mathcal{F}, \mu), (Y, \mathcal{G}, \nu)\)

Let \(K\) be the \(\pi\) system of measurable rectangles \(A \times B\) where \(A \in \mathcal{F}\) and \(B \in \mathcal{G}\). Explain why this is really a \(\pi\) system. Now let \(\mathcal{F} \times \mathcal{G}\) denote the smallest \(\sigma\) algebra which contains \(K\). Let

\[
\mathcal{P} = \left\{ A \in \mathcal{F} \times \mathcal{G} : \int_X \int_Y X_A \, d\nu \, d\mu = \int_Y \int_X X_A \, d\mu \, d\nu \right\}
\]

where both integrals make sense and are equal. Then show that \(\mathcal{P}\) is closed with respect to complements and countable disjoint unions. By Dynkin’s lemma, \(\mathcal{P} = \mathcal{F} \times \mathcal{G}\). Then define a measure \(\mu \times \nu\) as follows. For \(A \in \mathcal{F} \times \mathcal{G}\)

\[
\mu \times \nu (A) = \int_X \int_Y X_A \, d\nu \, d\mu
\]

Explain why this is a measure and why if \(f\) is \(\mathcal{F} \times \mathcal{G}\) measurable and nonnegative, then

\[
\int_{X \times Y} f \, d(\mu \times \nu) = \int_X \int_Y X_A \, d\nu \, d\mu = \int_Y \int_X X_A \, d\mu \, d\nu
\]

**Hint:** This is just a repeat of what I showed you in class except that it is easier because the measures are finite. Pay special attention to the way the monotone convergence theorem is used.

8. Let \((X, \mathcal{F}, \mu)\) be a regular measure space. For example, it could be \(\mathbb{R}^p\) with Lebesgue measure. Why do we care about a measure space being regular? This problem will show why. Suppose that closures of balls are compact as in the case of \(\mathbb{R}^p\).

(a) Let \(\mu(E) < \infty\). By regularity, there exists \(K \subseteq E \subseteq V\) where \(K\) is compact and \(V\) is open such that \(\mu(V \setminus K) < \varepsilon\). Show there exists \(W\) open such that \(K \subseteq W \subseteq V\) and \(W\) is compact. Now show there exists a function \(h\) such that \(h\) has values in \([0, 1]\), \(h(x) = 1\) for \(x \in K\), and \(h(x)\) equals 0 off \(W\). **Hint:** You might consider Problem 12 on Page 474.
(b) Show that
\[ \int |X_E - h| \, d\mu < \varepsilon \]

(c) Next suppose \( s = \sum_{i=1}^{n} c_i X_{E_i} \) is a nonnegative simple function where each \( \mu (E_i) < \infty \). Show there exists a continuous nonnegative function \( h \) which equals zero off some compact set such that
\[ \int |s - h| \, d\mu < \varepsilon \]

(d) Now suppose \( f \geq 0 \) and \( f \in L^1 (\Omega) \). Show that there exists \( h \geq 0 \) which is continuous and equals zero off a compact set such that
\[ \int |f - h| \, d\mu < \varepsilon \]

(e) If \( f \in L^1 (\Omega) \) with complex values, show the conclusion in the above part of this problem is the same.

9. Let \((\Omega, \mathcal{F}, \mu)\) be a measure space and suppose \( f, g : \Omega \to (-\infty, \infty] \) are measurable. Prove the sets
\[ \{ \omega : f(\omega) < g(\omega) \} \] and \[ \{ \omega : f(\omega) = g(\omega) \} \]
are measurable. **Hint:** The easy way to do this is to write
\[ \{ \omega : f(\omega) < g(\omega) \} = \bigcup_{r \in \mathbb{Q}} \{ f < r \} \cap \{ g > r \} . \]
Note that \( l (x, y) = x - y \) is not continuous on \((-\infty, \infty] \) so the obvious idea doesn’t work. Here \([g > r] \) signifies \( \{ \omega : g(\omega) > r \} \).

10. Let \( \{ f_n \} \) be a sequence of real or complex valued measurable functions. Let
\[ S = \{ \omega : \{ f_n(\omega) \} \text{ converges} \}. \]
Show \( S \) is measurable. **Hint:** You might try to exhibit the set where \( f_n \) converges in terms of countable unions and intersections using the definition of a Cauchy sequence.

11. Suppose \( u_n(t) \) is a differentiable function for \( t \in (a, b) \) and suppose that for \( t \in (a, b) \),
\[ |u_n(t)|, \ |u_n'(t)| < K_n \]
where \( \sum_{n=1}^{\infty} K_n < \infty \). Show
\[ (\sum_{n=1}^{\infty} u_n(t))' = \sum_{n=1}^{\infty} u_n'(t) . \]
**Hint:** This is an exercise in the use of the dominated convergence theorem and the mean value theorem.

12. Suppose \( \{ f_n \} \) is a sequence of nonnegative measurable functions defined on a measure space, \((\Omega, \mathcal{S}, \mu)\). Show that
\[ \int \sum_{k=1}^{\infty} f_k \, d\mu = \sum_{k=1}^{\infty} \int f_k \, d\mu . \]
**Hint:** Use the monotone convergence theorem along with the fact the integral is linear.
13. Explain why for each \( t > 0 \), \( x \to e^{-tx} \) is a function in \( L^1(\mathbb{R}) \) and
\[
\int_0^\infty e^{-tx} \, dx = \frac{1}{t}.
\]
Thus
\[
\int_0^R \frac{\sin(t)}{t} \, dt = \int_0^R \int_0^\infty \sin(t) e^{-tx} \, dx \, dt
\]
Now explain why you can change the order of integration in the above iterated integral. Then compute what you get. Next pass to a limit as \( R \to \infty \) and show
\[
\int_0^\infty \frac{\sin(t)}{t} \, dt = \frac{1}{2}\pi
\]
This is a very important integral. Note that the thing on the left is an improper integral. \( \sin(t)/t \) is not Lebesgue integrable because it is not absolutely integrable. That is
\[
\int_0^\infty \left| \frac{\sin t}{t} \right| \, dm = \infty
\]
It is important to understand that the Lebesgue theory of integration only applies to nonnegative functions and those which are absolutely integrable.

14. Show \( \lim_{n \to \infty} \frac{n}{2^n} \sum_{k=1}^n \frac{2^k}{k} = 2 \). This problem was shown to me by Shane Tang, a former student. It is a nice exercise in dominated convergence theorem if you massage it a little.
\[
\text{Hint:}
\]
\[
\frac{n}{2^n} \sum_{k=1}^n \frac{2^k}{k} = \sum_{k=1}^n \frac{2^{k-n}}{k} = \sum_{l=0}^{n-1} 2^{-l} \frac{n}{n-l} = \sum_{l=0}^{n-1} 2^{-l} \left( 1 + \frac{l}{n-l} \right) \leq \sum_{l=0}^{n-1} 2^{-l} (1 + l)
\]
15. Let the rational numbers in \([0,1]\) be \( \{r_k\}_{k=1}^\infty \) and define
\[
f_n(t) = \begin{cases} 1 & \text{if } t \in \{r_1, \ldots, r_n\} \\ 0 & \text{if } t \notin \{r_1, \ldots, r_n\} \end{cases}
\]
Show that \( \lim_{n \to \infty} f_n(t) = f(t) \) where \( f \) is one on the rational numbers and 0 on the irrational numbers. Explain why each \( f_n \) is Riemann integrable but \( f \) is not. However, each \( f_n \) is actually a simple function and its Lebesgue and Riemann integral is equal to 0. Apply the monotone convergence theorem to conclude that \( f \) is Lebesgue integrable and in fact, \( \int f \, dm = 0 \).

16. Give an example of a sequence of functions \( \{f_n\}, f_n \geq 0 \) and a function \( f \geq 0 \) such that \( f(x) = \lim \inf_{n \to \infty} f_n(x) \) but \( \int f \, dm \leq \lim \inf_{n \to \infty} \int f_n \, dm \) so you get strict inequality in Fatou’s lemma.

17. Let \( f \) be a nonnegative Riemann integrable function defined on \([a,b]\). Thus there is a unique number between all the upper sums and lower sums. First explain why, if \( a_i \geq 0 \),
\[
\int \sum_{i=1}^n a_i \chi_{[t_i, t_{i-1})}(t) \, dm = \sum_i a_i (t_i - t_{i-1})
\]
Explain why there exists an increasing sequence of Borel measurable functions \( \{g_n\} \) converging to a Borel measurable function \( g \), and a decreasing sequence of functions \( \{h_n\} \) which are also Borel measurable converging to a Borel measurable function \( h \) such that \( g_n \leq f \leq h_n \),
\[
\int g_n \, dm \text{ equals a lower sum} \\
\int h_n \, dm \text{ equals an upper sum}
\]
and \( \int (h - g) \, dm = 0 \). Explain why \( \{ x : f(x) \neq g(x) \} \) is a set of measure zero. Then explain why \( f \) is measurable and \( \int_a^b f(x) \, dx = \int f \, dm \) so that the Riemann integral gives the same answer as the Lebesgue integral.
Chapter 21

Measures From Positive Linear Functionals

Rudin does it this way and I really don’t know a better way to do it. In this chapter, we will only consider the measure space to be a complete separable metric space in which the measure of balls is finite. Rudin does this for an arbitrary locally compact Hausdorff space, but all examples of interest to me are metric spaces. In fact, the one of most interest is $\mathbb{R}^n$.

Definition 21.0.1 Let $\Omega$ be a Polish space (complete separable metric space). Define $C_c(\Omega)$ to be the functions which have complex values and compact support. This means $\text{spt}(f) = \{x \in \Omega : f(x) \neq 0\}$ is a compact set. Then $L : C_c(\Omega) \to \mathbb{C}$ is called a positive linear functional if it is linear and if, whenever $f \geq 0$, then $L(f) \geq 0$ also.

The following definition gives some notation.

Definition 21.0.2 If $K$ is a compact subset of an open set, $V$, then $K \prec \phi \prec V$ if

$$\phi \in C_c(V), \quad \phi(K) = \{1\}, \quad \phi(\Omega) \subseteq [0, 1],$$

where $\Omega$ denotes the whole topological space considered. Also for $\phi \in C_c(\Omega)$, $K \prec \phi$ if

$$\phi(\Omega) \subseteq [0, 1] \text{ and } \phi(K) = 1.$$

and $\phi \prec V$ if

$$\phi(\Omega) \subseteq [0, 1] \text{ and } \text{spt}(\phi) \subseteq V.$$

Now we need some lemmas.

Theorem 21.0.3 Let $H$ be a compact subset of an open set $U$ in a metric space having the property that the closures of balls are compact. Then there exists an open set $V$ such that

$$H \subseteq V \subseteq \tilde{V} \subseteq U$$

with $\tilde{V}$ compact. There also exists $\psi$ such that $H \prec f \prec V$, meaning that $f = 1$ on $H$ and $\text{spt}(f) \subseteq \tilde{V}$.

Proof: Consider $h \to \text{dist}(h, U^C)$. This continuous function achieves its minimum at some $h_0 \in H$ because $H$ is compact. Let $\delta \equiv \text{dist}(h_0, U^C)$. The distance is positive because $U^C$ is closed. Now $H \subseteq \cup_{h \in H} B(h, \delta)$. Since $H$ is compact, there are finitely many of these balls which cover $H$. Say $H \subseteq \cup_{i=1}^k B(h_i, \delta) \equiv V$. Then, since there are finitely many of these balls, $\tilde{V} = \cup_{i=1}^k \overline{B(h_i, \delta)}$ which is a compact set since it is a finite union of compact sets.
To obtain $f$, let

$$f(x) = \frac{\text{dist} (x, V^c)}{\text{dist} (x, V^c) + \text{dist} (x, H)}$$

Then $f(x) \leq 1$ and if $x \in H$, its distance to $V^c$ is positive and $\text{dist} (x, H) = 0$ so $f(x) = 1$. If $x \in V^c$, then its distance to $H$ is positive and so $f(x) = 0$. It is obviously continuous because the denominator is a continuous function and never vanishes. Thus $H < f < V$. ■

**Theorem 21.0.4 (Partition of unity)** Let $K$ be a compact subset of a Polish space in which the closures of balls are compact and suppose

$$K \subseteq V = \cup_{i=1}^n V_i, \text{ } V_i \text{ open.}$$

Then there exist $\psi_i < V_i$ with

$$\sum_{i=1}^n \psi_i(x) = 1$$

for all $x \in K$. If $H$ is a compact subset of $V_i$ for some $V_i$ there exists a partition of unity such that $\psi_i(x) = 1$ for all $x \in H$

**Proof**: Let $K_1 = K \setminus \cup_{i=2}^n V_i$. Thus $K_1$ is compact and $K_1 \subseteq V_1$. Let $K_1 \subseteq W_1 \subseteq \overline{W}_1 \subseteq V_1$ with $\overline{W}_1$ compact. To obtain $W_1$, use Theorem 21.0.3. to get $g$ such that $K_1 < f < V_1$ and let $W_1 \equiv \{x : f(x) \neq 0\}$. Thus $W_1, V_2, \cdots V_n$ covers $K$ and $\overline{W}_1 \subseteq V_1$. Let $K_2 = K \setminus (\cup_{i=2}^n V_i \cup W_1)$. Then $K_2$ is compact and $K_2 \subseteq V_2$. Let $K_2 \subseteq W_2 \subseteq \overline{W}_2 \subseteq V_2 \overline{W}_2$ compact. Continue this way finally obtaining $W_i, \cdots, W_n$, $K \subseteq W_1 \cup \cdots \cup W_n$, and $\overline{W}_i \subseteq V_i \overline{W}_i$ compact. Now let $W_1 \subseteq U_i \subseteq \overline{U}_i \subseteq V_i, \overline{U}_i$ compact.

By Theorem 21.0.3. let $U_i \prec \phi_i < V_i, \cup_{i=1}^n \overline{U}_i \prec \gamma < \cup_{i=1}^n U_i$. Define

$$\psi_i(x) = \begin{cases} \gamma(x)\phi_i(x) / \sum_{j=1}^n \phi_j(x) & \text{if } \sum_{j=1}^n \phi_j(x) \neq 0, \\ 0 & \text{if } \sum_{j=1}^n \phi_j(x) = 0. \end{cases}$$

If $x$ is such that $\sum_{j=1}^n \phi_j(x) = 0$, then $x \notin \cup_{i=1}^n \overline{U}_i$. Consequently $\gamma(y) = 0$ for all $y$ near $x$ and so $\psi_i(y) = 0$ for all $y$ near $x$. Hence $\psi_i$ is continuous at such a point. $x$. If $x \in K$, then $\gamma(x) = 1$ and so $\sum_{j=1}^n \psi_j(x) = 1$. Clearly $0 \leq \psi_i(x) \leq 1$ and $\text{spt} (\psi_i) \subseteq V_i$. As to the last claim, keep $V_i$ the same but replace $V_j$ with $V_j \equiv V_j \setminus H$. Now in the proof above, applied to this modified collection of open sets, if $j \neq i, \phi_j(x) = 0$ whenever $x \in H$. Therefore, $\psi_i(x) = 1$ on $H$. ■

Now with this preparation, here is the main result called the Riesz representation theorem for positive linear functionals.

**Theorem 21.0.5 (Riesz representation theorem)** Let $(\Omega, \tau)$ be a Polish space for which the closures of the balls are compact and let $L$ be a positive linear functional on $C_c(\Omega)$. Then there exists a $\sigma$ algebra $\mathcal{S}$ containing the Borel sets and a unique measure $\mu$, defined on $\mathcal{S}$, such that

$$\mu$$ is complete,

$$\mu(K) < \infty \text{ for all } K \text{ compact},$$

$$\mu(F) = \sup \{\mu(K) : K \subseteq F, \text{ } K \text{ compact}\},$$

for all $F \in \mathcal{S},$

$$\mu(F) = \inf \{\mu(V) : V \supseteq F, \text{ } V \text{ open}\}$$

for all $F \in \mathcal{S},$ and

$$\int f \, d\mu = LF \text{ for all } f \in C_c(\Omega).$$

(21.1)
The plan is to define an outer measure and then to show that it, together with the \( \sigma \) algebra of sets measurable in the sense of Caratheodory, satisfies the conclusions of the theorem. Always, \( K \) will be a compact set and \( V \) will be an open set.

**Definition 21.0.6** \( \mu(V) \equiv \sup \{ Lf : f \prec V \} \) for \( V \) open, \( \mu(\emptyset) = 0 \). \( \mu(E) \equiv \inf \{ \mu(V) : V \supseteq E \} \) for arbitrary sets \( E \).

**Lemma 21.0.7** \( \mu \) is a well-defined outer measure.

**Proof:** First it is necessary to verify \( \mu \) is well defined because there are two descriptions of it on open sets. Suppose then that \( \mu_1(V) = \inf \{ \mu(U) : U \supseteq V \text{ and } U \text{ is open} \} \). It is required to verify that \( \mu_1(V) = \mu(V) \) where \( \mu \) is given as \( \sup \{ Lf : f \prec V \} \). If \( U \supseteq V \), then \( \mu(U) \geq \mu(V) \) directly from the definition. Hence from the definition of \( \mu_1 \), it follows \( \mu_1(V) \geq \mu(V) \). On the other hand, \( V \supseteq V \) and so \( \mu_1(V) \leq \mu(V) \). This verifies \( \mu \) is well defined. \( \blacksquare \)

It remains to show that \( \mu \) is an outer measure. Let \( V = \cup_{i=1}^{\infty} V_i \) and let \( f \prec V \). Then \( \operatorname{spt}(f) \subseteq \cup_{i=1}^{n} V_i \) for some \( n \). Let \( \psi_i \prec V_i \), \( \sum_{i=1}^{n} \psi_i = 1 \) on \( \operatorname{spt}(f) \).

\[
Lf = \sum_{i=1}^{n} L(f \psi_i) \leq \sum_{i=1}^{n} \mu(V_i) \leq \sum_{i=1}^{\infty} \mu(V_i).
\]

Hence

\[
\mu(V) \leq \sum_{i=1}^{\infty} \mu(V_i)
\]

since \( f \prec V \) is arbitrary.

Now let \( E = \cup_{i=1}^{\infty} E_i \). Is \( \mu(E) \leq \sum_{i=1}^{\infty} \mu(E_i) \)? Without loss of generality, it can be assumed \( \mu(E_i) < \infty \) for each \( i \) since if not so, there is nothing to prove. Let \( V_i \supseteq E_i \) with \( \mu(E_i) + \varepsilon 2^{-i} > \mu(V_i) \).

\[
\mu(E) \leq \mu(\cup_{i=1}^{\infty} V_i) \leq \sum_{i=1}^{\infty} \mu(V_i) \leq \varepsilon + \sum_{i=1}^{\infty} \mu(E_i).
\]

Since \( \varepsilon \) was arbitrary, \( \mu(E) \leq \sum_{i=1}^{\infty} \mu(E_i) \) which proves the lemma. \( \blacksquare \)

**Lemma 21.0.8** Let \( K \) be compact, \( g \geq 0 \), \( g \in C_c(\Omega) \), and \( g = 1 \) on \( K \). Then \( \mu(K) \leq Lg \). Also \( \mu(K) < \infty \) whenever \( K \) is compact.

**Proof:** Let \( \alpha \in (0, 1) \) and \( V_\alpha = \{ x : g(x) > \alpha \} \) so \( V_\alpha \supseteq K \) and let \( h \prec V_\alpha \).

Then \( h \leq 1 \) on \( V_\alpha \) while \( g \alpha^{-1} \geq 1 \) on \( V_\alpha \) and so \( g \alpha^{-1} \geq h \) which implies \( L(g \alpha^{-1}) \geq Lh \) and that therefore, since \( L \) is linear,

\[
Lg \geq \alpha Lh.
\]

Since \( h \prec V_\alpha \) is arbitrary, and \( K \subseteq V_\alpha \),

\[
Lg \geq \alpha \mu(V_\alpha) \geq \alpha \mu(K).
\]

Letting \( \alpha \uparrow 1 \) yields \( Lg \geq \mu(K) \). This proves the first part of the lemma. The second assertion follows from this and Theorem 21.0.3. If \( K \) is given, let

\[
K \prec g \prec \Omega
\]

and so from what was just shown, \( \mu(K) \leq Lg < \infty \). \( \blacksquare \)
Lemma 21.0.9 If $A$ and $B$ are disjoint subsets of $\Omega$, with $\text{dist} \,(A,B) > 0$ then $\mu(A \cup B) = \mu(A) + \mu(B)$.

**Proof:** There is nothing to show if $\mu(A \cup B) = \infty$. Thus we can let $\delta \equiv \text{dist} \,(A,B) > 0$. Then let $U_1 \equiv \bigcup_{a \in A} B \left(a, \frac{\delta}{4}\right); V_1 \equiv \bigcup_{b \in B} B \left(b, \frac{\delta}{4}\right)$. It follows that these two open sets have empty intersection. Also, there exists $W \supseteq A \cup B$ such that $\mu(W) - \varepsilon < \mu(A \cup B)$. Let $U \equiv U_1 \cap W, V \equiv V_1 \cap W$. Then

$$\mu(A \cup B) + \varepsilon > \mu(W) \geq \mu(U \cup V)$$

Now let $f < U, g < V$ such that $Lf + \varepsilon > \mu(U), Lg + \varepsilon > \mu(V)$. Then

$$\mu(U \cup V) \geq L (f + g) = L(f) + L(g) > \mu(U) - \varepsilon + (\mu(V) - \varepsilon) \geq \mu(A) + \mu(B) - 2\varepsilon$$

It follows that

$$\mu(A \cup B) + \varepsilon > \mu(A) + \mu(B) - 2\varepsilon$$

and since $\varepsilon$ is arbitrary, $\mu(A \cup B) \geq \mu(A) + \mu(B) \geq \mu(A \cup B)$.

It follows from Theorem 21.0.5 that the measurable sets $\mathcal{S}$ contains the Borel $\sigma$-algebra $\mathcal{B}(\Omega)$. Since closures of balls are compact, it follows from Lemma 21.0.9 that $\mu$ is finite on every ball. Corollary 21.0.10 implies that $\mu$ is regular for every $E$ a Borel set. That is,

$$\mu(E) = \sup \{\mu(K) : K \subseteq E\},$$

$$\mu(E) = \inf \{\mu(V) : V \supseteq E\}$$

In particular, $\mu$ is inner regular on every open set $V$. This is obtained immediately. In fact the same thing holds for any $F \in \mathcal{S}$ in place of $E$ in the above. The second of the two follows immediately from the definition of $\mu$. It remains to verify the first. In doing so, first assume that $\mu(F)$ is contained in a closed ball $B$. Let $V \supseteq (B \setminus F)$ such that

$$\mu(V) < \mu(B \setminus F) + \varepsilon$$

Then $\mu(V \setminus (B \setminus F)) + \mu(B \setminus F) = \mu(V) < \mu(B \setminus F) + \varepsilon$ and so $\mu(V \setminus (B \setminus F)) < \varepsilon$. Now consider $V \cap F$. This is a closed subset of $F$. To see that it is closed, note that $V \cap F = V \cap B$ which is a closed set. Why is this so? It is clear that $V \cap F \subseteq V \cap B$. Now if $x \in V \cap B$, then since $V \supseteq (B \setminus F)$, it follows that $x \in V \subseteq B \cap F$ and so either $x \in B \cap F$ which doesn’t occur, or $x \in F$ and so this must be the case. Hence, $V \cap B$ is a closed, hence compact subset of $F$. Now

$$\mu(F \setminus (V \cap B)) = \mu(F \cap (V \cup B \cap F)) \leq \mu(V \setminus B) \leq \mu(V \setminus (B \setminus F)) < \varepsilon$$

It follows that $\mu(F) < \mu(V \cap B) + \varepsilon$ which shows inner regularity in case $F$ is contained in some closed ball $B$. If this is not the case, let $B_n$ be a sequence of closed balls having increasing radii and let $F_n = B_n \cap F$. If $l < \mu(F_n)$, it follows that $\mu(F_n) > l$ for all large enough $n$. Then picking one of these, it follows from what was just shown that there is a compact set $K \subseteq F_n$ such that also $\mu(K) > l$.

Thus $\mathcal{S}$ contains the Borel sets and $\mu$ is inner regular on all sets of $\mathcal{S}$.

It remains to show $\mu$ satisfies 21.0.8.

Lemma 21.0.10 If $f$ is a positive measure $\mu = Lf$ for all $f \in C_c(\Omega)$.

**Proof:** Let $f \in C_c(\Omega)$, $f$ real-valued, and suppose $f(\Omega) \subseteq [a,b]$. Choose $t_0 < a$ and let $t_0 < t_1 < \cdots < t_n = b, t_i - t_{i-1} < \varepsilon$. Let

$$E_i = f^{-1}((t_{i-1}, t_i]] \cap \text{spt}(f).$$

(21.4)
Note that $\bigcup_{i=1}^n E_i$ is a closed set, and in fact

$$\bigcup_{i=1}^n E_i = \text{spt}(f)$$

(21.5)

since $\Omega = \bigcup_{i=1}^n f^{-1}((t_{i-1}, t_i])$. Let $V_i \supseteq E_i$, $V_i$ is open and let $V_i$ satisfy

$$f(\omega) < t_i + \varepsilon \text{ for all } \omega \in V_i,$$  

(21.6)

$$\mu(V_i \setminus E_i) < \varepsilon/n.$$

By Theorem 21.0.4 there exists $h_i \in C_c(\Omega)$ such that

$$h_i \prec V_i, \quad \sum_{i=1}^n h_i(\omega) = 1 \text{ on } \text{spt}(f).$$

Now note that for each $i$,

$$f(\omega)h_i(\omega) \leq h_i(\omega)(t_i + \varepsilon).$$

(If $\omega \in V_i$, this follows from 21.0.6. If $\omega \notin V_i$ both sides equal 0.) Therefore,

$$Lf = L(\sum_{i=1}^n fh_i) \leq L(\sum_{i=1}^n h_i(t_i + \varepsilon))$$

$$= \sum_{i=1}^n (t_i + \varepsilon) L(h_i)$$

$$= \sum_{i=1}^n (|t_0| + t_i + \varepsilon)L(h_i) - |t_0|L\left(\sum_{i=1}^n h_i\right).$$

Now note that $|t_0| + t_i + \varepsilon \geq 0$ and so from the definition of $\mu$ and Lemma 21.0.8, this is no larger than

$$\sum_{i=1}^n (|t_0| + t_i + \varepsilon)\mu(V_i) - |t_0|\mu(\text{spt}(f))$$

$$\leq \sum_{i=1}^n (|t_0| + t_i + \varepsilon)\left(\mu(E_i) + \varepsilon/n\right) - |t_0|\mu(\text{spt}(f))$$

$$\leq |t_0|\sum_{i=1}^n \mu(E_i) + |t_0|\varepsilon + \sum_{i=1}^n t_i\mu(E_i) + \varepsilon(|t_0| + |b|)$$

$$+ \sum_{i=1}^n t_i \frac{\varepsilon}{n} + \varepsilon \sum_{i=1}^n \mu(E_i) + \varepsilon^2 - |t_0|\mu(\text{spt}(f)).$$

From 21.5 and 21.6, the first and last terms cancel. Therefore this is no larger than

$$(2|t_0| + |b| + \mu(\text{spt}(f)) + \varepsilon)\varepsilon$$

$$+ \sum_{i=1}^n t_{i-1}\mu(E_i) + \varepsilon\mu(\text{spt}(f)) + \sum_{i=1}^n (|t_0| + |b|) \frac{\varepsilon}{n}$$

$$\leq \int f d\mu + (2|t_0| + |b| + 2\mu(\text{spt}(f)))\varepsilon + (|t_0| + |b|) \varepsilon$$

Since $\varepsilon > 0$ is arbitrary,

$$Lf \leq \int f d\mu \quad (21.7)$$
for all \( f \in C_c(\Omega) \), \( f \) real. Hence equality holds because \( L(-f) \leq -\int f \, du \) so \( L(f) \geq \int f \, du \). Thus \( Lf = \int f \, du \) for all \( f \in C_c(\Omega) \). Just apply the result for real functions to the real and imaginary parts of \( f \). This proves the Lemma.

This gives the existence part of the Riesz representation theorem.

It only remains to prove uniqueness. Suppose both \( \mu_1 \) and \( \mu_2 \) are measures on \( S \) satisfying the conclusions of the theorem. Then if \( K \) is compact and \( V \supseteq K \), let \( K \prec f \prec V \). Then

\[
\mu_1(K) \leq \int f \, d\mu_1 = Lf = \int f \, d\mu_2 \leq \mu_2(V).
\]

Thus \( \mu_1(K) \leq \mu_2(K) \) for all \( K \). Similarly, the inequality can be reversed and so it follows the two measures are equal on compact sets. By the assumption of inner regularity on open sets, the two measures are also equal on all open sets. By outer regularity, they are equal on all sets of \( S \).

The regularity of this measure is very significant. Here is a useful lemma.

**Lemma 21.0.11** Suppose \( f : \Omega \rightarrow [0, \infty) \) is measurable where \( \mu \) is a regular measure as in the above theorem having the measure of any ball finite and the closures of balls compact. Then there is a set of measure zero \( N \) and an increasing sequence of functions \( \{h_n\}, h_n : \Omega \rightarrow [0, \infty) \) each in \( C_c(\Omega) \) such that for all \( \omega \in \Omega \setminus N, h_n(\omega) \rightarrow f(\omega) \). Also, for \( \omega \notin N, h_n(\omega) \leq f(\omega) \).

**Proof:** Consider \( f_n(\omega) \equiv X_{B_n}(\omega) \min(f(\omega), n) \) where \( B_n \) is a ball centered at \( \omega_0 \) which has radius \( n \). Thus \( f_n(\omega) \) is an increasing sequence and converges to \( f(\omega) \) for each \( \omega \). Also by Corollary 21.1.6, there exists a simple function \( s_n \) such that

\[
s_n(\omega) \leq f_n(\omega), \quad \sup_{\omega \in \Omega} |f_n(\omega) - s_n(\omega)| < \frac{1}{2^n}
\]

Let

\[
s_n(\omega) = \sum_{k=1}^{m_n} c_k^n \chi_{E_k^n}(\omega), \quad c_k^n > 0
\]

Then it must be the case that \( \mu(E_k^n) < \infty \) because \( \int f_n \, d\mu < \infty \).

By regularity, there exists a compact set \( K_n^k \) and an open set \( V_k^n \) such that

\[
K_n^k \subseteq E_k^n \subseteq V_k^n, \quad \sum_{k=1}^{m_n} \mu(V_k^n \setminus K_n^k) < \frac{1}{2^n}
\]

Now let \( K_n^k \prec \psi_k^n \prec V_k^n \) and let

\[
h_n(\omega) \equiv \sum_{k=1}^{m_n} c_k^n \psi_k^n(\omega)
\]

Thus for \( N_n = \bigcup_{k=1}^{m_n} V_k^n \setminus K_n^k \), it follows \( \mu(N_n) < 1/2^n \) and

\[
\sup_{\omega \notin N_n} |f_n(\omega) - h_n(\omega)| < \frac{1}{2^n}
\]

If \( h_n(\omega) \) fails to converge to \( f(\omega) \), then \( \omega \) must be in infinitely many of the \( N_n \). That is,

\[
\omega \in \cap_{n=1}^{\infty} \bigcup_{k \geq n} N_k \equiv N
\]

However, this set \( N \) is contained in

\[
\bigcup_{k=n}^{\infty} N_k, \quad \mu(\bigcup_{k=n}^{\infty} N_k) \leq \sum_{k=n}^{\infty} \mu(N_k) < \frac{1}{2^{n-1}}
\]

and so \( \mu(N) = 0 \). If \( \omega \) is not in \( N \), then eventually \( \omega \) fails to be in \( N_n \) and also \( \omega \in B_n \) so \( h_n(\omega) = s_n(\omega) \) for all \( n \) large enough. Now \( f_n(\omega) \rightarrow f(\omega) \) and \( |s_n(\omega) - f_n(\omega)| < 1/2^n \) so also \( s_n(\omega) = h_n(\omega) \rightarrow f(\omega) \).

Note that each \( N_k \) is an open set and so, \( N \) is a Borel set. Thus the above lemma leads to the following corollary.


**Corollary 21.0.12** Let $f$ be measurable. Then there exists a Borel measurable function $g$ and a Borel set of measure zero $N$ such that $f(\omega) = g(\omega)$ for all $\omega \notin N$. In fact, off $N$, $f(\omega) = \lim_{n \to \infty} h_n(\omega)$ where $h_n$ is continuous and $|h_n(\omega)| \leq |f(\omega)|$.

**Proof:** Apply the above lemma to the positive and negative parts of the real and imaginary parts of $f$. Let $N$ be the union of the exceptional Borel sets which result. Thus, $fX_{NC}$ is the limit of a sequence $h_nX_{NC}$ where $h_n$ is continuous, $|h_n(\omega)| \leq |f(\omega)|$ for $\omega \notin N$. Thus $h_nX_{NC}$ is Borel and it follows that $fX_{NC}$ is Borel measurable. Let $g = fX_{NC}$. ■

When you have a measure space $(\Omega, \mathcal{F}, \mu)$ in which $\Omega$ is a topological space and $\mu$ is regular, complete, and finite on compact sets, then $\mu$ is called a Radon measure. The above Riesz representation theorem gives an important example of Radon measures, those which come from a positive linear functional on a metric space in which the closed balls are compact.

### 21.1 Lebesgue Measure On $\mathbb{R}^n$, Fubini’s Theorem

In the above representation theorem, let

$$Lf = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(x_1, \cdots, x_n) \, dx_i_1 \cdots dx_i_n \quad (21.8)$$

where $(i_1, \cdots, i_n)$ is some permutation of $(1, \cdots, n)$ and $f \in C_c(\Omega)$. These are the familiar iterated improper Riemann integrals. Thus

$$\int_{\mathbb{R}} f(x_1, \cdots, x_n) \, dx_i_1 = \lim_{r \to \infty} \int_{-r}^{r} f(x_1, \cdots, x_n) \, dx_i_1 = \int_{-a}^{a} f(x_1, \cdots, x_n) \, dx_i_1$$

for all $a$ sufficiently large since $f$ has compact support. The measure which results is called $m_n$. It is $n$ dimensional Lebesgue measure and has all the regularity properties mentioned in the Riesz representation theorem above. In particular, $m_1$ refers to $\mathbb{R}$ and

$$\int f(x) \, dx = \int f(x) \, dm_1$$

for $f$ continuous with compact support. I will continue using the $dx$ notation in the context of iterated integrals instead of something like the following which is arguably more consistent with the above notation but more cumbersome to think about and write.

$$\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f dm_1(x_1) \cdots dm_1(x_n)$$

**Lemma 21.1.1** Let $E$ be Borel. Then if $f \in C_c(\mathbb{R}^n)$ is nonnegative,

$$\int_{\mathbb{R}^n} fX_E \, dm_n = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} fX_E dx_i_1 \cdots dx_i_n \quad (*)$$

with each iterated integral making sense. Also,

$$\int_{\mathbb{R}^n} X_E \, dm_n = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} X_E dx_i_1 \cdots dx_i_n \quad (**)$$

with all the iterated integrals making sense.

**Proof:** Let $K$ denote sets of the form $\prod_{i=1}^{n} (a_i, b_i)$. Then it is clearly a $\pi$ system. Also, since every open set is the countable union of sets of this form, $\sigma(K) \supseteq \mathcal{B}(\mathbb{R}^n)$, the Borel sets because it contains the open sets. Then $*$ holds for any set in $K$ thanks to an approximation like that used in determining the Lebesgue measure of a box and the monotone or dominated convergence theorem. Now let $G$ denote the measurable sets for which $*$ holds. It clearly contains $K$ and is closed with
respect to complements and countable disjoint unions and so by Dynkin’s lemma, Lemma 21.1.1, \( G \supseteq \sigma (\mathcal{K}) \supseteq \mathcal{B}(\mathbb{R}^n) \). The reason is it closed with respect to complements is that if \( E \in G \),

\[
\int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} f \chi_{E} \, dx_{i_1} \cdots dx_{i_n} + \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} f \chi_{E^c} \, dx_{i_1} \cdots dx_{i_n}
\]

\[
= \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} f \, dx_{i_1} \cdots dx_{i_n} = \int_{\mathbb{R}^n} f \, dm_n = \int_{\mathbb{R}^n} f \chi_{E} \, dm_n + \int_{\mathbb{R}^n} f \chi_{E^c} \, dm_n
\]

Then cancelling \( \int_{\mathbb{R}^n} f \chi_{E} \, dx_{i_1} \cdots dx_{i_n} \) and \( \int_{\mathbb{R}^n} f \chi_{E^c} \, dm_n \) yields

\[
\int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} f \chi_{E^c} \, dx_{i_1} \cdots dx_{i_n} = \int_{\mathbb{R}^n} f \chi_{E} \, dm_n
\]

The claim about countable disjoint unions is obvious.

As to the last claim, let \( f_k \) be an increasing sequence which converges to 1. Then by the monotone convergence theorem, ** follows from

\[
\int_{\mathbb{R}^n} \chi_{E} \, dm_n = \lim_{k \to \infty} \int_{\mathbb{R}^n} f_k \chi_{E} \, dm_n = \lim_{k \to \infty} \int_{\mathbb{R}^n} f_k \chi_{E} \, dx_{i_1} \cdots dx_{i_n}
\]

\[
= \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} \chi_{E} \, dx_{i_1} \cdots dx_{i_n}
\]

The last equality follows from repeated iterations of the monotone convergence theorem for each successive iterated integral, described more in the next major theorem. ■

The following is often called Fubini’s theorem.

** Theorem 21.1.2 ** Let \( f(x) \geq 0 \) and let \( f \) be Borel measurable. Then

\[
\int_{\mathbb{R}^n} f \, dm_n = \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} f(x_1, \cdots, x_n) \, dx_{i_1} \cdots dx_{i_n}
\]

(*)

where \((i_1, \cdots, i_n)\) is a permutation of \((1, 2, \cdots, n)\) and all iterated integrals make sense.

** Proof:** Let \( \{s_k\} \) be an increasing sequence of Borel measurable simple functions converging to \( f \) pointwise. Then \( * \) holds with \( f \) replaced with \( s_k \) by Lemma 21.1.1. Now apply the monotone convergence theorem to obtain the result. In the right side you must apply it to the successive iterated integrals to yield that the iterated integrals all make sense and in the limit yield the desired iterated integral. To illustrate, note that

\[
\int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} \chi_{E} \, dx_{i_1} \cdots dx_{i_n}
\]

is measurable with respect to one dimensional Lebesgue measure by the above lemma. This allows moving the limit inside the first integral from the left in

\[
\int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} s_k(x_1, \cdots, x_n) \, dx_{i_1} \cdots dx_{i_{n-1}}
\]

Then the same reasoning allows one to move it inside the next integral from the left and so forth, each time preserving the one dimensional Lebesgue measurability. Thus the above is increasing to

\[
= \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} \lim_{k \to \infty} s_k(x_1, \cdots, x_n) \, dx_{i_1} \cdots dx_{i_{n-1}}
\]

There is an easy corollary which is very important.
Corollary 21.1.3 Lebesgue measure is translation invariant. That is, if \( F \) is Lebesgue measurable, then \( m_n(F) = m_n(x + F) \).

**Proof:** Let \( K \) denote sets of the form \( \prod_{i=1}^{n} I_i \) where \( I_i \) is an interval. Then if \( H \) is one of these sets, Fubini’s theorem shows that \( m_n(H) = m_n(x + H) \) since in each case, the measure is the product of the lengths of the sides of a box. Now let \( R_k = \prod_{j=1}^{n} [-k, k] \). Let \( G \) denote those Borel sets \( E \) such that \( m_n(E \cap R_k) = m_n(x + E \cap R_k) \). As just noted, \( K \subseteq G \) because the intersection is just something in \( K \) if \( E \in K \). Now if \( E \in G \), then

\[
m_n(R_k \cap E^C) = m_n(R_k) - m_n(R_k \cap E) = m_n(x + R_k) - m_n(x + R_k \cap E) = m_n(x + R_k \cap E^C)
\]

If \( \{E_i\} \) are disjoint and each in \( G \) then

\[
m_n(R_k \cap \bigcup_{j=1}^{\infty} E_j) = \sum_{j=1}^{\infty} m_n(R_k \cap E_j) = \sum_{j=1}^{\infty} m_n(x + R_k \cap E_j) = m_n(x + R_k \cap (\bigcup_{j=1}^{\infty} E_j))
\]

Thus \( G \) is closed with respect to countable disjoint unions and complements. Therefore, by Dynkin’s lemma, \( G \supseteq \sigma(K) \) but \( \alpha(K) \) obviously includes all open sets so it contains the Borel sets. Since it is a subset of the Borel sets, it is equal to the Borel sets. Now let \( k \to \infty \) to conclude that \( m_n(E) = m_n(x + E) \) for all \( E \) Borel.

Now by regularity, if \( F \) is any measurable set, there exists \( E \) a countable union of compact sets and \( G \) a countable intersection of open sets such that \( E \subseteq F \subseteq G \) and

\[
m_n(E) = m_n(F) = m_n(G)
\]

Then \( m_n \) is translation invariant for both \( E, G \) because these are Borel sets. Therefore, it is translation invariant for \( F \) also. Specifically,

\[
m_n(x + F) \leq m_n(x + G) = m_n(G) = m_n(F) = m_n(E) = m_n(x + E) \leq m_n(x + F)
\]

Therefore, all inequalities are actually equalities and this shows the desired result. \( \blacksquare \)

In the following proposition, \( \|\cdot\| \) refers to a norm on \( \mathbb{R}^n \), not necessarily the usual one. Actually something better will be proved because it is more convenient to do so.

Proposition 21.1.4 Let \( \alpha(n) \) denote the \( m_n \) measure of the ball \( B(0,1) = \{x \in \mathbb{R}^n : \|x\| < 1\} \). Then

\[
m_n(B(x, r)) = \alpha(n) r^n.
\]

**Proof:** More generally, let \( G \) denote those Borel sets with the following property. Letting \( R_p = \prod_{i=1}^{n} [-p, p] \), for all \( r > 0 \),

\[
m_n(r(E \cap R_p)) = r^n m_n(E \cap R_p).
\]

Let \( K \) denote those sets of the form \( \prod_{k=1}^{n} I_k \) where \( I_k \) is an interval. Then if \( U \equiv \prod_{k=1}^{n} [a_k, b_k] \) is one of these sets, you would have \( U \cap R_p = \prod_{k=1}^{n} [a_k, b_k] \) or else \( \emptyset \) in which case, there is nothing to prove, and so the condition holds because

\[
m_n(r(U \cap R_p)) = \prod_{k=1}^{n} r(b_k - a_k) = r^n m_n(U \cap R_p)
\]
The situation is the same if the intervals are open or half open or if they are a mixture of various kinds of intervals. Thus $\mathcal{K} \subseteq \mathcal{G}$. Now if you have disjoint sets of $\mathcal{G} \{ E_i \}$, then the sets $rE_i$ are also disjoint. Then

$$m_n ( r ( R_p \cap \bigcup_i E_i) ) = m_n ( \bigcup_i r ( R_p \cap E_i) ) = \sum_i m_n ( r ( R_p \cap E_i) ) = r^n m_n ( R_p \cap E_i)$$

and so $\mathcal{G}$ is closed with respect to countable disjoint unions. Now say $E \in \mathcal{G}$. What about $E^C$?

$$m_n ( r ( R_p \cap E^C) ) = m_n ( r R_p ) - m_n ( r ( R_p \cap E) ) = r^n ( m_n ( R_p ) - m_n ( R_p \cap E) ) = r^n m_n ( R_p \cap E^C)$$

and so $\mathcal{G}$ is closed with respect to complements also. By Dynkin’s lemma, it follows that $\mathcal{G} = \sigma ( \mathcal{K} ) = \mathcal{B} ( \mathbb{R}^n )$. Thus whenever $E$ is Borel,

$$m_n ( r ( E \cap R_p ) ) = r^n m_n ( E \cap R_p )$$

Now let $p \to \infty$ to obtain the desired result when $E$ is Borel. Thus in particular,

$$m_n ( B ( x, r ) ) = m_n ( B ( 0, r ) ) = r^n m_n ( B ( 0, 1 ) ) = \alpha ( n ) r^n .$$

21.2 The Besicovitch Covering Theorem

There are other covering theorems such as the Vitali covering theorem, but this theorem depends on adjusting the size of the balls. The Besicovitch covering theorem doesn’t do this. Instead, it estimates the number of intersections. This is why it has no problem in dealing with arbitrary Radon measures. It is an amazing result but even though this is the case, it really only depends on simple properties of normed linear spaces and so fits well with linear algebra.

The first fundamental observation is found in the following lemma which holds for the context illustrated by the following picture. This picture is drawn such that the balls come from the usual Euclidean norm, but the norm could be any norm on $\mathbb{R}^n$. Also, it is not particularly important whether the balls are open or closed. They are just balls which may or may not contain points on their boundary.

The idea is to consider balls $B_i$ which intersect a given ball $B$ such that $B$ contains no center of any $B_i$ and no $B_i$ contains the center of another $B_j$. There are two cases to consider, the case where the balls have large radii and the case where the balls have small radii.

**Intersections with big balls**

**Lemma 21.2.1** Let the balls $B_\alpha, B_x, B_y$ be as shown, having radii $r, r_x, r_y$ respectively. Suppose the centers of $B_x$ and $B_y$ are not both in any of the balls shown, and suppose $r_y \geq r_x \geq \alpha r$ where $\alpha$ is a number larger than 1. Also let $P_x = a + r \frac{x-a}{\|x-a\|} \alpha r$ with $P_y$ being defined similarly. Then it follows that $\| P_x - P_y \| \geq \frac{2-\alpha}{\alpha+1} r$. There exists a constant $L ( n, \alpha )$ depending on $\alpha$ and the dimension, such that if $B_1, \cdots , B_m$ are all balls such that any pair are in the same situation relative to $B_\alpha$ as $B_x$, and $B_y$, then $m \leq L ( n, \alpha )$. 

21.2. THE BESICOVITCH COVERING THEOREM

Proof: From the definition,

\[ ||P_x - P_y|| = r \left( \frac{|x - a|}{|x - a|} - \frac{|y - a|}{|y - a|} \right) \]

\[ = r \left( \frac{|x - a| |y - a| - (y - a)|x - a|}{|x - a| |y - a|} \right) \]

\[ = r \frac{|y - a| (x - y) + (y - a)(|y - a| - |x - a|)}{|x - a| |y - a|} \]

\[ \geq r \frac{|x - y|}{|x - a|} - r \frac{|y - a| |y - a| - |x - a|}{|x - a| |y - a|} \]

\[ = r \frac{|x - y|}{|x - a|} - r \frac{|y - a|}{|x - a|} \frac{|y - a| - |x - a|}{|y - a|} \]  \hspace{1cm} (21.9)

There are two cases. First suppose that \(|y - a| - |x - a| \geq 0\). Then the above

\[ = r \frac{|x - y|}{|x - a|} - r \frac{|y - a|}{|x - a|} + r. \]

From the assumptions, \(|x - y| \geq r_y\) and also \(|y - a| \leq r + r_y\). Hence the above

\[ \geq r \left( \frac{r_y}{|x - a|} - \frac{r + r_y}{|x - a|} + 1 \right) \geq r \left( 1 - \frac{r}{|x - a|} \right) \]

\[ \geq r \left( 1 - \frac{r}{r_x} \right) \geq r \left( 1 - \frac{1}{\alpha} \right) = r \left( \frac{\alpha - 1}{\alpha} \right) \geq \frac{r \alpha - 1}{\alpha + 1}. \]

The other case is that \(|y - a| - |x - a| < 0\) in 21.3. Then in this case 21.3 equals

\[ r \left( \frac{|x - y|}{|x - a|} - \frac{1}{|x - a|} (|y - a| - |x - a|) \right) \]

\[ = r \left( \frac{|x - y|}{|x - a|} - \frac{1}{|x - a|} (|x - a| - |y - a|) \right) \]

Then since \(|x - a| \leq r + r_x, |x - y| \geq r_y, |y - a| \geq r_y, \)

\[ \frac{r}{r_x + r} (r_y - (r + r_x) + r_y) \geq \frac{r}{r_x + r} (r_y - (r + r_y) + r_y) \]

\[ \geq \frac{r}{r_x + r} (r_y - r) \geq \frac{r}{r_x + r} (r_x - r) \geq \frac{r}{r_x + 1/\alpha r_x} \left( r_x - \frac{1}{\alpha} r_x \right) \]

\[ = \frac{r}{1 + (1/\alpha)} (1 - 1/\alpha) = \frac{\alpha - 1}{\alpha + 1} r \]

This proves the estimate between \(P_x\) and \(P_y\).

Finally, in the case of the balls \(B_i\) having centers at \(x_i\), then as above, let \(P_{x_i} = a + r \frac{x_i - a}{||x_i - a||}\).

Then \((P_{x_i} - a) r^{-1}\) is on the unit sphere having center 0. Furthermore,

\[ ||(P_{x_i} - a) r^{-1} - (P_y - a) r^{-1}|| = r^{-1} ||P_{x_i} - P_y|| \geq r^{-1} \frac{\alpha - 1}{\alpha + 1} = \frac{\alpha - 1}{\alpha + 1}. \]

How many points on the unit sphere can be pairwise this far apart? This set is compact and so there exists a \(\frac{1}{3} \left( \frac{\alpha - 1}{\alpha + 1} \right)\) net having \(L(n, \alpha)\) points. Thus \(m\) cannot be any larger than \(L(n, \alpha)\) because if it were, then by the pigeon hole principal, two of the points \((P_{x_i} - a) r^{-1}\) would lie in a single ball \(B \left( p, \frac{1}{3} \left( \frac{\alpha - 1}{\alpha + 1} \right) \right)\) so they could not be \(\frac{\alpha - 1}{\alpha + 1}\) apart. \(\blacksquare\)
The above lemma has to do with balls which are relatively large intersecting a given ball. Next is a lemma which has to do with relatively small balls intersecting a given ball. Note that in the statement of this lemma, the radii are smaller than \( r \) in contrast to the above lemma in which the radii of the balls are larger than \( r \). In the application of this lemma, we will have \( \gamma = 4/3 \) and \( \beta = 1/3 \). These constants will come from a construction, while \( \alpha \) is just something larger than 1 which we will take here to equal 10.

### Intersections with small but comparable balls

**Lemma 21.2.2** Let \( B \) be a ball having radius \( r \) and suppose \( B \) has nonempty intersection with the balls \( B_1, \ldots, B_m \) having radii \( r_1, \ldots, r_m \) respectively, and as before, no \( B_i \) has the center of any other and the centers of the \( B_i \) are not contained in \( B \). Suppose \( \alpha, \gamma > 1 \) and the \( r_i \) are comparable with \( r \) in the sense that

\[
\frac{1}{\gamma} r \leq r_i \leq \alpha r.
\]

Let \( B'_i \) have the same center as \( B_i \) with radius equal to \( \beta r_i \) for some \( \beta < 1 \). If the \( B'_i \) are disjoint, then there exists a constant \( M(n, \alpha, \beta, \gamma) \) such that \( m \leq M(n, \alpha, \beta, \gamma) \). Letting \( \alpha = 10, \beta = 1/3, \gamma = 4/3 \), it follows that \( m \leq 60^n \).

**Proof:** Let the volume of a ball of radius \( r \) be given by \( \alpha(n) r^n \) where \( \alpha(n) \) depends on the norm used and on the dimension \( n \) as indicated. The idea is to enlarge \( B \), till it swallows all the \( B'_i \). Then, since they are disjoint and their radii are not too small, there can’t be too many of them.

This can be done for a single \( B'_i \) by enlarging the radius of \( B \) to \( r + r_i + r'_i \).

\[ \begin{align*}
B_i & \quad B'_i
\end{align*} \]

Then to get all the \( B_i \), you would just enlarge the radius of \( B \) to \( r + \alpha r + \beta \alpha r = (1 + \alpha + \alpha \beta) r \). Then, using the inequality which makes \( r_i \) comparable to \( r \), it follows that

\[
\sum_{i=1}^{m} \alpha(n) \left( \frac{\beta}{\gamma} r \right)^n \leq \sum_{i=1}^{m} \alpha(n) (\beta r_i)^n \leq \alpha(n) (1 + \alpha + \alpha \beta)^n r^n
\]

Therefore,

\[
m \left( \frac{\beta}{\gamma} \right)^n \leq (1 + \alpha + \alpha \beta)^n
\]

and so \( m \leq (1 + \alpha + \alpha \beta)^n \left( \frac{\beta}{\gamma} \right)^{-n} \equiv M(n, \alpha, \beta, \gamma) \).

From now on, let \( \alpha = 10 \) and let \( \beta = 1/3 \) and \( \gamma = 4/3 \). Then

\[
M(n, \alpha, \beta, \gamma) \leq \left( \frac{172}{3} \right)^n \leq 60^n
\]

Thus \( m \leq 60^n \). ■

The next lemma gives a construction which yields balls which are comparable as described in the above lemma. \( r(B) \) will denote the radius of the ball \( B \).

**A construction of a sequence of balls**
Lemma 21.2.3 Let $\mathcal{F}$ be a nonempty set of nonempty balls in $\mathbb{R}^n$ with
\[ \sup \{\text{diam}(B) : B \in \mathcal{F} \} \leq D < \infty \]
and let $A$ denote the set of centers of these balls. Suppose $A$ is bounded. Define a sequence of balls from $\mathcal{F}$, $\{B_j\}_{j=1}^J$ where $J \leq \infty$ such that
\[ r(B_1) \geq \frac{3}{4} \sup \{r(B) : B \in \mathcal{F}\} \]
and if
\[ A_m \equiv A \setminus (\bigcup_{i=1}^m B_i) \neq \emptyset, \]
then $B_{m+1} \in \mathcal{F}$ is chosen with center in $A_m$ such that
\[ r_{m+1} \equiv r(B_{m+1}) \geq \frac{3}{4} \sup \{r(B, a) : B \in \mathcal{F}, a \in A_m\}. \]
Then letting $B_j = B(a_j, r_j)$, this sequence satisfies
\[ r(B_k) \leq \frac{4}{3} r(B_j) \text{ for } j < k, \]
\[ \{B(a_j, r_j/3)\}_{j=1}^J \text{ are disjoint}, \]
\[ A \subseteq \bigcup_{i=1}^J B_i. \]

Proof: Consider Lemma 21.2.3. First note the sets $A_m$ form a decreasing sequence. Thus from the definition of $B_j$, for $j < k$,
\[ r(B_k) \leq \sup \{r(B, a) : B \in \mathcal{F}, a \in A_{k-1}\} \leq \sup \{r(B, a) : B \in \mathcal{F}, a \in A_{j-1}\} \leq \frac{4}{3} r(B_j) \]
because the construction gave
\[ r(B_j) \geq \frac{3}{4} \sup \{r(B, a) : B \in \mathcal{F}, a \in A_{j-1}\} \]

Next consider Lemma 21.2.3. If $x \in B(a_j, r_j/3) \cap B(a_i, r_i/3)$ where these balls are two which are chosen by the above scheme such that $j > i$, then from what was just shown
\[ \|a_j - a_i\| \leq \|a_j - x\| + \|x - a_i\| \leq r_j + r_i \leq \frac{4}{3} + \frac{1}{3} \leq \frac{7}{9} r_i < r_i \]
and this contradicts the construction because $a_i$ is not covered by $B(a_i, r_i)$.

Finally consider Lemma 21.2.3. If $x \in B(a_j, r_j/3) \cap B(a_i, r_i/3)$ where these balls are two which are chosen by the above scheme such that $j > i$, then from what was just shown
\[ \|a_j - a_i\| \leq \|a_j - x\| + \|x - a_i\| \leq r_j + r_i \leq \frac{4}{3} + \frac{1}{3} \leq \frac{7}{9} r_i < r_i \]
and this contradicts the construction because $a_i$ is not covered by $B(a_i, r_i)$.

As explained above, in this sequence of balls from the above lemma, if $j < k$
\[ \frac{3}{4} r(B_k) \leq r(B_j) \]
Then there are two cases to consider,
\[ r(B_j) \geq 10 r(B_k), \ r(B_j) \leq 10 r(B_k) \]
In the first case, we use Lemma 21.2.3 to estimate the number of intersections of $B_k$ with $B_j$ for $j < k$. In the second case, we use Lemma 21.2.3 to estimate the number of intersections of $B_k$ with $B_j$ for $j < k$.

Now here is the Besicovitch covering theorem.
There exists a constant $N_n$, depending only on $n$ with the following property. If $\mathcal{F}$ is any collection of nonempty balls in $\mathbb{R}^n$ with
\[
\sup \{ \text{diam} (B) : B \in \mathcal{F} \} < D < \infty
\]
and if $A$ is the set of centers of the balls in $\mathcal{F}$, then there exist subsets of $\mathcal{F}$, $\mathcal{H}_1, \cdots, \mathcal{H}_{N_n}$, such that each $\mathcal{H}_i$ is a countable collection of disjoint balls from $\mathcal{F}$ (possibly empty) and $A \subseteq \bigcup_{i=1}^{N_n} \bigcup \{ B : B \in \mathcal{H}_i \}$.

**Proof:** To begin with, suppose $A$ is bounded. Let $L(n,10)$ be the constant of Lemma 21.2.4 and let $M_n = L(n,10)+60^n+1$. Define the following sequence of subsets of $\mathcal{F}$, $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_{M_n}$. Referring to the sequence $\{ B_k \}$ just considered, let $B_1 \in \mathcal{G}_1$ and if $B_1, \cdots, B_m$ have been assigned, each to a $\mathcal{G}_i$, place $B_{m+1}$ in the first $\mathcal{G}_j$ such that $B_{m+1}$ intersects no set already in $\mathcal{G}_j$. The existence of such a $j$ follows from Lemmas 21.2.4 and 21.2.2. Here is why. $B_{m+1}$ can intersect at most $L(n,10)$ sets of $\{ B_1, \cdots, B_m \}$ which have radii at least as large as $10B_{m+1}$ thanks to Lemma 21.2.4. It can intersect at most $60^n$ sets of $\{ B_1, \cdots, B_m \}$ which have radius smaller than $10B_{m+1}$ thanks to Lemma 21.2.2. Thus each $\mathcal{G}_j$ consists of disjoint sets of $\mathcal{F}$ and the set of centers is covered by the union of these $\mathcal{G}_j$. This proves the theorem in case the set of centers is bounded.

Now let $R_1 = B(0,5D)$ and if $R_m$ has been chosen, let $R_{m+1} = B(0,(m+1)5D) \setminus R_m$.

Thus, if $|k-m| \geq 2$, no ball from $\mathcal{F}$ having nonempty intersection with $R_m$ can intersect any ball from $\mathcal{F}$ which has nonempty intersection with $R_k$. This is because all these balls have radius less than $D$. Now let $A_m = A \cap R_m$ and apply the above result for a bounded set of centers to those balls of $\mathcal{F}$ which intersect $R_m$ to obtain sets of disjoint balls $\mathcal{G}_1(R_m), \mathcal{G}_2(R_m), \cdots, \mathcal{G}_{M_n}(R_m)$ covering $A_m$. Then simply define $\mathcal{G}'_j = \bigcup_{k=1}^{\infty} \mathcal{G}_j(R_{2k})$, $\mathcal{G}_j = \bigcup_{k=1}^{\infty} \mathcal{G}_j(R_{2k-1})$. Let $N_n = 2M_n$ and
\[
\{ \mathcal{H}_1, \cdots, \mathcal{H}_{N_n} \} \equiv \{ \mathcal{G}'_1, \cdots, \mathcal{G}'_{M_n}, \mathcal{G}_1, \cdots, \mathcal{G}_{M_n} \}
\]
Note that the balls in $\mathcal{G}'_j$ are disjoint. This is because those in $\mathcal{G}_j(R_{2k})$ are disjoint and if you consider any ball in $\mathcal{G}_j(R_{2k})$, it cannot intersect a ball of $\mathcal{G}_j(R_{2m})$ for $m \neq k$ because $|2k-2m| \geq 2$. Similar considerations apply to the balls of $\mathcal{G}_j$.

Now let $f: \mathbb{R}^n \to \mathbb{R}^n$ be a Lipschitz function. This means there is a constant $K$ such that
\[
\| f(x) - f(y) \| \leq K \| x - y \|
\]
where $\| \cdot \|$ denotes a norm on $\mathbb{R}^n$. For example, $f(x)$ could equal $Ax$ where $A$ is an $n \times n$ matrix. In this case,
\[
\| Ax - Ay \| \leq \| A \| \| x - y \|
\]
with $\| A \|$ the operator norm. Then the following proposition is a fundamental result which says that a Lipschitz map of a measurable set is a measurable set. This is a remarkable result because it is not even true if $f$ is only continuous. For example, see Problem 8 on Page 183.

**Proposition 21.2.5** Let $f: \mathbb{R}^n \to \mathbb{R}^n$ be Lipschitz continuous with Lipschitz constant $K$. Then if $F$ is a Lebesgue measurable set, then so is $f(F)$. Also if $N$ is a set of measure zero, then $f(N)$ is a set of measure zero.

**Proof:** Consider the second claim first. Let $V$ be an open set, $m_n(V) < \varepsilon$, and $V \supseteq K$. For each point $x$ of $K$, there is a closed ball $B(x,r) \subseteq V$ such that also $r < 1$. Let $\mathcal{F}$ denote this collection of balls and let $\{ \mathcal{H}_1, \cdots, \mathcal{H}_{N_n} \}$ be the finite set each of which consists of countably many disjoint balls from $\mathcal{F}$ whose union includes all of $N$. Denote by $f(\mathcal{H}_k)$ the set of $f(B)$ where $B \in \mathcal{H}_k$. Each is compact because $f$ is continuous and $B$ is closed. Thus $f(N)$ is contained in the union of the $f(\mathcal{H}_k)$. It follows that
\[
m_n(f(N)) \leq \sum_{k=1}^{N_n} \sum_{B \in \mathcal{H}_k} m_n(f(B))
\]
Here the bar on the measure denotes the outer measure determined by \( m_n \). Now \( f(B) \) is contained in a ball of the form \( B(f(x), Kr) \) where \( B = B(x, r) \) and so, by Proposition 21.1.3, \( m_n(f(B)) \leq K^n r^n \alpha(n) \), and also \( \alpha(n) r^n = m_n(B) \). Therefore, the above reduces to

\[
\overline{m}_n(f(N)) \leq \sum_{k=1}^{N_n} \sum_{B \in \mathcal{R}_k} m_n(f(B)) \leq K^n \sum_{k=1}^{N_n} \sum_{B \in \mathcal{R}_k} m_n(B)
\]

\[
\leq K^n N_n m_n(V) < K^n N_n \varepsilon
\]

Since \( \varepsilon \) is arbitrary, this implies that \( \overline{m}_n(f(N)) = 0 \) and so \( f(N) \) is Lebesgue measurable and in fact is a set of measure zero. This is by completeness of Lebesgue measure. See Proposition [\textit{MM5}] and completeness of Lebesgue measure.

For the first part, suppose \( F \) is Lebesgue measurable. Then there exists \( H \subseteq F \) such that \( H \) is the countable union of compact sets and also that \( m_n(H) = m_n(F) \). Thus \( H \) is measurable because it is the union of measurable sets. Say \( m_n(F) < \infty \) first. Then \( m_n(F \setminus H) = 0 \) and so \( f(F \setminus H) \) is a measurable set of measure zero. Then \( f(F) = f(F \setminus H) \cup f(H) \). The second is obviously measurable because it is a countable union of compact sets, the continuous image of a compact set being compact. It follows that \( f(F) \) is also measurable. In general, consider \( F_k \equiv F \cap [-k, k]^n \). Then \( f(F_k) \) is measurable and now \( f(F) = \cup_k f(F_k) \) so it is also measurable. ■

### 21.3 Change Of Variables, Linear Map

It was shown above that if \( A \) is an \( n \times n \) matrix, then \( AE \) is Lebesgue measurable whenever \( E \) is.

**Lemma 21.3.1** Let \( A \) be an \( n \times n \) matrix and let \( E \) be Lebesgue measurable. Then

\[
m_n(A(E)) = |\det(A)| m_n(E)
\]

**Proof:** This is to be shown first for elementary matrices. To begin with, let \( A \) be the elementary matrix which involves adding the \( i^{th} \) row of the identity to the \( j^{th} \) row. Thus

\[
A(x_1, \ldots, x_n) = (x_1, \ldots, x_i + x_j, \ldots, x_n)
\]

Consider what it does to \( B \equiv \prod_{k=1}^{n} [a_k, b_k] \). The \( j^{th} \) variable now goes from \( a_j + x_i \) to \( b_j + x_i \). Thus, by Fubini’s theorem

\[
\int_{\mathbb{R}^n} A(B) \, dm_n = \int_{\mathbb{R}^n} \chi_{A(B)}(x_1, \ldots, x_n) \, dm_n
\]

\[
= \int \cdots \int \chi_{A(B)}(x_1, \ldots, x_n) \, dx_j \, dx_i \, dx_1 \cdots dx_{j-1} \, dx_{j+1} \cdots dx_n
\]

the integration taken in an order such that the first two two are with respect to \( x_i \) and \( x_j \). Then from what was just observed, this reduces to

\[
\int \chi_{[a_1, b_1]}(x_1) \cdots \int \chi_{[a_n, b_n]}(x_n) \int \chi_{[a_i, b_i]}(x_i) \int \chi_{[a_j + x_i, b_j + x_i]}(x_j) \, dx_j \, dx_i \, dx_1 \cdots dx_n
\]

That inner integral is still \( (b_j - a_j) \) and so the whole thing reduces to \( \prod_k (b_k - a_k) = m_n(B) \). The determinant of this elementary matrix is 1 because it is either upper triangular or lower triangular with all ones down the main diagonal. Thus in this case, \( m_n(A(B)) = |\det(A)| m_n(B) \). In case \( A \) is the elementary matrix which involves multiplying the \( i^{th} \) row by \( \alpha \neq 0 \), \( |\det(A)| = |\alpha| \) and \( A(B) \) is also a box which changes only the interval corresponding to \( x_i \), making it \( |\alpha| \) times as long. Thus \( m_n(A(B)) = |\alpha| m_n(B) = |\det(A)| m_n(B) \). The remaining kind of elementary matrix \( A \) involves switching two rows of \( I \) and it takes \( B \) to \( \overline{B} \) where \( \overline{B} \) is the Cartesian product of the same intervals as in \( B \) but the intervals now occur with respect to different variables. Thus the \( m_n(A(B)) = m_n(B) \).

Also, \( |\det(A)| = | - \det(I) | = 1 \) and so again \( m_n(A(B)) = |\det(A)| m_n(B) \).
Letting $A$ be one of these elementary matrices, let $G$ denote the Borel sets $E$ such that
\[ m_n (A (E \cap R_m)) = |\det (A)| \, m_n (E \cap R_m), \quad R_m = [-m, m]^n \quad (21.16) \]

Then if $\mathcal{K}$ consists of sets of the form $\prod_{k=1}^n [a_k, b_k],$ the above shows that $\mathcal{K} \subseteq \mathcal{G}$. Also it is clear that $\mathcal{G}$ is closed with respect to countable disjoint unions and complements. Therefore, $\mathcal{G} \supseteq \sigma (\mathcal{K})$. But $\sigma (\mathcal{K}) \supseteq \mathcal{B} (\mathbb{R}^n)$ because $\sigma (\mathcal{K})$ clearly contains every open set. Now let $m \to \infty$ in (21.14) to obtain the desired result.

Next suppose that $F$ is a Lebesgue measurable set, $m_n (F) < \infty$. Then by inner regularity, there is a set $E \subseteq F$ such that $E$ is the countable union of compact sets, hence Borel, and $m_n (F) = m_n (E)$. Thus $m_n (F \setminus E) = 0,$ since $m_n (F) = 0$. By Proposition 21.2, 21.3,
\[ m_n (A (F)) \leq m_n (A (E)) + m_n (A (F \setminus E)) = |\det (A)| \, m_n (E) \leq |\det (A)| \, m_n (F) \quad (*) \]

Thus, also
\[ m_n (F) = m_n (A^{-1} (AF)) \leq |\det (A^{-1})| \, m_n (A (F)) \]
and so
\[ |\det (A)| \, m_n (F) \leq m_n (A (F)) \]
This with * shows that $|\det (A)| \, m_n (F) = m_n (A (F)).$ In the general case,
\[ |\det (A)| \, m_n (F \cap R_m) = m_n (A (F \cap R_m)) \]
Now let $m \to \infty$ to get the desired result.

It follows from iterating this result that if $A$ is any product of elementary matrices, $E_1 \cdots E_r,$ then for $F$ Lebesgue measurable,
\[ m_n (A (F)) = m_n (E_1 \cdots E_r (F)) = |\det (E_1)| \, m_n (E_2 \cdots E_r (F)) \cdots = \prod_{i=1}^r |\det (E_i)| \, m_n (F) = |\det (A)| \, m_n (F) \]

If $A$ is not the product of elementary matrices, then it has rank less than $n$ and so there are elementary matrices, $E_1, \cdots, E_r$ such that
\[ E_1 \cdots E_r A = B \]
where $B$ is in row reduced echelon form and has at least one row of zeros. Thus, if $F$ is a Lebesgue measurable set, then
\[ 0 = m_n (B (F)) = m_n (E_1 \cdots E_r A (F)) = \prod_{i=1}^r |\det (E_i)| \, m_n (A (F)) \]
The first equality comes from Fubini’s theorem and Proposition 21.4, 21.5.
\[ B (F) \subseteq \{(x_1, \cdots, x_n) \in \mathbb{R}^n : x_n = 0\} \equiv C, \quad a \text{ Borel set} \]

Therefore, by Fubini’s theorem,
\[ m_n (B (F)) \leq m_n (C) = \int \cdots \int \chi_C (x_1, \cdots, x_n) \, dx_1 \cdots dx_n = 0 \]
Thus, even in this case, $m_n (A (F)) = |\det (A)| \, m_n (F).$ □
21.4 Vitali Coverings

There is another covering theorem which may also be referred to as the Besicovitch covering theorem. As before, the balls can be taken with respect to any norm on $\mathbb{R}^n$. At first, the balls will be closed but this assumption will be removed.

**Definition 21.4.1** A collection of balls, $\mathcal{F}$ covers a set, $E$ in the sense of Vitali if whenever $x \in E$ and $\varepsilon > 0$, there exists a ball $B \in \mathcal{F}$ whose center is $x$ having diameter less than $\varepsilon$.

I will give a proof of the following theorem.

**Theorem 21.4.2** Let $\mu$ be a Radon measure on $\mathbb{R}^n$ and let $E$ be a set with $\overline{\mu}(E) < \infty$. Where $\overline{\mu}$ is the outer measure determined by $\mu$. Suppose $\mathcal{F}$ is a collection of closed balls which cover $E$ in the sense of Vitali. Then there exists a sequence of disjoint balls, $\{B_i\} \subseteq \mathcal{F}$ such that

$$\overline{\mu}(E \setminus \bigcup_{j=1}^{\infty} B_j) = 0.$$  

**Proof:** Let $N_n$ be the constant of the Besicovitch covering theorem. Choose $r > 0$ such that

$$(1 - r)^{-1} \left(1 - \frac{1}{2N_n + 2}\right) \equiv \lambda < 1.$$

If $\overline{\mu}(E) = 0$, there is nothing to prove so assume $\overline{\mu}(E) > 0$. Let $U_1$ be an open set containing $E$ with $(1 - r) \mu(U_1) < \overline{\mu}(E)$ and $2\overline{\mu}(E) > \mu(U_1)$, and let $\mathcal{F}_1$ be those sets of $\mathcal{F}$ which are contained in $U_1$ whose centers are in $E$. Thus $\mathcal{F}_1$ is also a Vitali cover of $E$. Now by the Besicovitch covering theorem proved earlier, there exist balls $B_i$ of $\mathcal{F}_1$ such that

$$E \subseteq \bigcup_{i=1}^{N_n} \{B : B \in \mathcal{G}_i\}$$

where $\mathcal{G}_i$ consists of a collection of disjoint balls of $\mathcal{F}_1$. Therefore,

$$\overline{\mu}(E) \leq \sum_{i=1}^{N_n} \sum_{B \in \mathcal{G}_i} \mu(B)$$

and so, for some $i \leq N_n$,

$$(N_n + 1) \sum_{B \in \mathcal{G}_i} \mu(B) > \overline{\mu}(E).$$

It follows there exists a finite set of balls of $\mathcal{G}_i$, $\{B_1, \ldots, B_m\}$ such that

$$(N_n + 1) \sum_{i=1}^m \mu(B_i) > \overline{\mu}(E)$$

and so

$$(2N_n + 2) \sum_{i=1}^m \mu(B_i) > 2\overline{\mu}(E) > \mu(U_1).$$

Since $2\overline{\mu}(E) \geq \mu(U_1)$, $\overline{\mu}(E)$ implies

$$\frac{\mu(U_1)}{2N_n + 2} \leq \frac{2\overline{\mu}(E)}{2N_n + 2} = \frac{\overline{\mu}(E)}{N_n + 1} < \sum_{i=1}^m \mu(B_i).$$

Also $U_1$ was chosen such that $(1 - r) \mu(U_1) < \overline{\mu}(E)$, and so

$$\lambda\overline{\mu}(E) \geq \lambda(1 - r) \mu(U_1) = \left(1 - \frac{1}{2N_n + 2}\right) \mu(U_1)$$
Let 

\[ \geq \mu(U_1) - \sum_{i=1}^{m} \mu(B_i) = \mu(U_1) - \mu(\bigcup_{j=1}^{m} B_j) \]

\[ = \mu(U_1 \setminus \bigcup_{j=1}^{m} B_j) \geq \mathfrak{m}(E \setminus \bigcup_{j=1}^{m} B_j). \]

Since the balls are closed, you can consider the sets of \( F \) which have empty intersection with \( \bigcup_{j=1}^{m} B_j \) and this new collection of sets will be a Vitali cover of \( E \setminus \bigcup_{j=1}^{m} B_j \). Letting this collection of balls play the role of \( F \) in the above argument and letting \( E \setminus \bigcup_{j=1}^{m} B_j \) play the role of \( E \), repeat the above argument and obtain disjoint sets of \( F \),

\[ \{B_{m_1+1}, \ldots, B_{m_2}\}, \]

such that

\[ \lambda^k \mathfrak{m}(E \setminus \bigcup_{j=1}^{m} B_j) > \mathfrak{m}\left(\left(E \setminus \bigcup_{j=1}^{m_1} B_j\right) \setminus \bigcup_{j=m_1+1}^{m_2} B_j\right) = \mathfrak{m}(E \setminus \bigcup_{j=1}^{m_2} B_j), \]

and so

\[ \lambda^k \mathfrak{m}(E) > \mathfrak{m}(E \setminus \bigcup_{j=1}^{m} B_j). \]

Continuing in this way, yields a sequence of disjoint balls \( \{B_i\} \) contained in \( F \) and

\[ \mathfrak{m}(E \setminus \bigcup_{j=1}^{m} B_j) \leq \mathfrak{m}(E \setminus \bigcup_{j=1}^{m_k} B_j) < \lambda^k \mathfrak{m}(E) \]

for all \( k \). Therefore, \( \mathfrak{m}(E \setminus \bigcup_{j=1}^{\infty} B_j) = 0 \) and this proves the Theorem. \( \square \)

It is not necessary to assume \( \mathfrak{m}(E) < \infty \).

**Corollary 21.4.3** Let \( \mu \) be a Radon measure on \( \mathbb{R}^n \). Letting \( \mathfrak{m} \) be the outer measure determined by \( \mu \), suppose \( F \) is a collection of closed balls which cover \( E \) in the sense of Vitali. Then there exists a sequence of disjoint balls, \( \{B_i\} \subseteq F \) such that

\[ \mathfrak{m}(E \setminus \bigcup_{j=1}^{\infty} B_j) = 0. \]

**Proof:** Since \( \mu \) is a Radon measure it is finite on compact sets. Therefore, there are at most countably many numbers, \( \{b_i\}_{i=1}^{\infty} \) such that \( \mu(\partial B(0,b_i)) > 0 \). It follows there exists an increasing sequence of positive numbers, \( \{r_i\}_{i=1}^{\infty} \) such that \( \lim_{i \to \infty} r_i = \infty \) and \( \mu(\partial B(0,r_i)) = 0 \). Now let

\[ D_1 = \{x : ||x|| < r_1\},D_2 = \{x : r_1 < ||x|| < r_2\}, \]

\[ \cdots , D_m = \{x : r_{m-1} < ||x|| < r_m\}, \ldots . \]

Let \( F_m \) denote those closed balls of \( F \) which are contained in \( D_m \). Then letting \( E_m = E \cap D_m \), \( F_m \) is a Vitali cover of \( E_m \), \( \mathfrak{m}(E_m) < \infty \), and so by Theorem \( 21.4.3 \) there exists a countable collection of balls from \( F_m \) \( \{B_j \}_{j=1}^{\infty} \), such that \( \mathfrak{m}(E_m \setminus \bigcup_{j=1}^{\infty} B_j) = 0 \). Then consider the countable collection of balls, \( \{B_j \}_{j,m=1}^{\infty} \).

\[ \mathfrak{m}(E \setminus \bigcup_{j=1}^{\infty} B_j) \leq \mathfrak{m}(\bigcup_{j=1}^{\infty} \partial B(0,r_j)) + \]

\[ + \sum_{m=1}^{\infty} \mathfrak{m}(E_m \setminus \bigcup_{j=1}^{\infty} B_j) = 0 \quad \square \]

You don’t need to assume the balls are closed. In fact, the balls can be open, closed or anything in between and the same conclusion can be drawn.

**Corollary 21.4.4** Let \( \mu \) be a Radon measure on \( \mathbb{R}^n \). Letting \( \mathfrak{m} \) be the outer measure determined by \( \mu \), suppose \( F \) is a collection of balls which cover \( E \) in the sense of Vitali, open closed or neither. Then there exists a sequence of disjoint balls, \( \{B_i\} \subseteq F \) such that

\[ \mathfrak{m}(E \setminus \bigcup_{j=1}^{\infty} B_j) = 0. \]
21.5. Change Of Variables

Here is an interesting proposition.

Proposition 21.5.1 Let \( f : U \to \mathbb{R}^n \) be differentiable on the open set \( U \subseteq \mathbb{R}^n \). Then if \( F \) is a Lebesgue measurable set, then so is \( f(K) \). Also if \( N \subseteq U \) is a set of measure zero, then \( f(N) \) is a set of measure zero.

Proof: Consider the second claim first. Let \( N \) be a set of measure zero and let

\[ N_k = \{ x \in N : \|Df(x)\| \leq k \} \]

There is an open set \( V \supseteq N_k \) such that \( m_n(V) < \varepsilon \). For each \( x \in N_k \), there is a ball \( B_x \) centered at \( x \) with radius \( r_x < 1 \) such that \( B_x \subseteq V \) and for \( y \in B_x \),

\[ f(y) \in f(x) + Df(x)(0, r_x) + B(0, \varepsilon r_x) \]

Thus

\[ f(B_x) \subseteq f(x) + B(0, (\|Df(x)\| + \varepsilon) r_x) \]

\[ \overline{m}_n(f(B_x)) \leq (k + \varepsilon)^n m_n(B(x, r_x)) \]

By the Besicovitch covering theorem, there are balls of this sort such that

\[ N_k \subseteq \bigcup_{j=1}^{M_n} B \in \mathcal{G}_j \]

where \( \mathcal{G}_j \) is a countable disjoint collection of these balls. Thus,

\[ \overline{m}_n(f(N_k)) \leq \sum_{j=1}^{M_n} \sum_{B \in \mathcal{G}_j} \overline{m}_n(f(B)) \leq (k + \varepsilon)^n \sum_{j=1}^{M_n} m_n(B) \]

\[ \leq (k + \varepsilon)^n M_n m_n(V) \leq \varepsilon (k + \varepsilon)^n M_n \]

Since \( \varepsilon \) is arbitrary, it follows that \( \overline{m}_n(f(N_k)) = 0 \) and so in fact \( f(N_k) \) is measurable and has \( m_n \) measure zero. See Proposition 11.4.1. Now let \( k \to \infty \) to conclude that \( m_n(f(N)) = 0 \). Now the conclusion is shown the same as in Proposition 21.4.3. You exploit inner regularity, and what was just shown, to obtain the conclusion that \( f(F) \) is measurable if \( F \) is.

Recall Lemma 11.4.3 which was based on the Brouwer fixed point theorem. The version of use here is stated below. In what follows, \(|\cdot|\) is the usual Euclidean norm on \( \mathbb{R}^n \).

Lemma 21.5.2 Let \( h \) be continuous and map \( \overline{B}(0, r) \subseteq \mathbb{R}^n \) to \( \mathbb{R}^n \). Suppose that for all \( x \in \overline{B}(0, r) \),

\[ |h(x) - x| < \varepsilon r \]

Then it follows that

\[ h(B(0, r)) \supseteq B(0, (1 - \varepsilon) r) \]
Now the Besicovitch covering theorem for a Vitali cover is used to give an easy treatment of the change of variables for multiple integrals. This will be based on the following lemma.

**Lemma 21.5.3** Let $h : U \to \mathbb{R}^n$ where $U$ is an open bounded set and suppose that $Dh (x)$ exists and is invertible and continuous on $U$ and that $h$ is one to one on $U$. Let $A \subseteq U$ be a Lebesgue measurable set. Then

$$m_n (h (A)) = \int_A |\det Dh (x)| \, dm_n$$

**Proof:** In what follows, $\varepsilon$ will be a small positive number. Let $A$ be a Lebesgue measurable subset of $U$ and

$$A \subseteq \{ x \in U : |\det Dh (x)| < k \} \quad (**)$$

let $V$ be an open set containing $A$ such that $m_n (V \setminus A) < \varepsilon$. We can also assume that $m_n (h (V) \setminus h (A)) < \varepsilon$. The reason for this is as follows. There exists $G \supseteq A$ such that $G$ is the countable intersection of nested open sets which each contain $A$ and $m_n (G \setminus A) = 0$. Say $G = \cap_i V_i$. Then since $h$ is one to one,

$$m_n (h (G)) - m_n (h (A)) = m_n (h (G) \setminus h (A)) = m_n (h (G \setminus A)) = 0$$

Then

$$m_n (h (A)) = m_n (h (G)) = \lim_{m \to \infty} m_n (h (V_m))$$

so eventually, for large enough $m$,

$$m_n (h (A)) + \varepsilon > m_n (h (V_m))$$

Then for $x \in A$,

$$h (x + v) = h (x) + Dh (x) v + o (v) \quad (21.18)$$

$$((h (x + v) - h (x)) - Dh (x) v) = o (v)$$

$$Dh (x)^{-1} (h (x + v) - h (x)) - v = o (v)$$

Thus, for all $v$ small enough, $|v| \leq r$,

$$\left| Dh (x)^{-1} (h (x + v) - h (x)) - v \right| < \varepsilon r$$

It follows from Lemma 21.5.3 that for each $x \in A$, there is a ball $B (x, r_x) \subseteq V$

$$Dh (x)^{-1} (h (x + B (0, \lambda r_x)) - h (x)) \supseteq B (0, (1 - \varepsilon) r)$$

$$h (x + B (0, \lambda r_x)) - h (x) \supseteq Dh (x) B (0, (1 - \varepsilon) r)$$

$$h (B (x, \lambda r_x)) \supseteq Dh (x) B (h (x), (1 - \varepsilon) r)$$

this holding for all $r_x$ small enough. Here $\lambda > 1$ is arbitrary. Thus

$$m_n (h (B (x, \lambda r_x))) \geq |\det (Dh (x))| (1 - \varepsilon)^n m_n (B (x, r_x))$$

Letting $\lambda \downarrow 1$, and using that $h$ maps sets of measure zero to sets of measure zero,

$$m_n (h (B (x, r_x))) \geq |\det (Dh (x))| (1 - \varepsilon)^n m_n (B (x, r_x))$$

As explained earlier, even if you have a general Radon measure, you could assume that $r_x$ is such $\{ y : |x - y| = r_x \}$ has measure zero, although in the case of Lebesgue measure, this can be shown for all $r_x$. It follows from Proposition 21.5.3 for example.
Also from (4.11.8), we can assume that \( r_x \) is small enough that

\[
 h(B(x, r_x)) - h(x) \subseteq Dh(x)B(0, (1 + \varepsilon)r_x) \\
h(B(x, r_x)) \subseteq Dh(x)B(h(x), (1 + \varepsilon)r_x)
\]

and so

\[
m_n(h(B(x, r_x))) \leq |\det Dh(x)|(1 + \varepsilon)^n m_n(B(x, r_x))
\]

Thus, whenever \( r_x \) is small enough,

\[
(1 - \varepsilon)^n |\det Dh(x)| m_n(B(x, r_x)) \\
\leq m_n(h(B(x, r_x))) \\
\leq (1 + \varepsilon)^n |\det Dh(x)| m_n(B(x, r_x))
\]

At this point, assume a little more on \( h \). Assume it is actually \( C^1(U) \). Then by making \( r_x \) smaller if necessary, it can be assumed that for \( y \in B(x, r_x) \),

\[
||\det Dh(x)| - |\det Dh(y)|| < \varepsilon
\]

This is a Vitali cover for \( A \). Therefore, there is a countable disjoint sequence of these balls \( \{B_i(x_i, r_i)\} \) such that \( m_n(A \setminus \bigcup_i B_i) = 0 \). Then since \( h \) is one to one,

\[
m_n(h(A)) \leq \sum_{i=1}^{\infty} m_n(h(B_i)) \leq m_n(h(V)) < m_n(h(A)) + \varepsilon
\]

Now the following chain of inequalities follow from the above.

\[
\int_A |Dh(x)|(1 - \varepsilon)^n \, dm_n \leq \sum_{i=1}^{\infty} \int_{B_i} |Dh(x)|(1 - \varepsilon)^n \, dm_n
\]

\[
\leq \sum_{i=1}^{\infty} \int_{B_i} (|Dh(x)| + \varepsilon)(1 - \varepsilon)^n \, dm_n
\]

\[
\leq \sum_{i=1}^{\infty} m_n(h(B_i)) + \varepsilon(1 - \varepsilon)^n m_n(U) \leq m_n(h(V)) + \varepsilon(1 - \varepsilon)^n m_n(U)
\]

\[
\leq m_n(h(A)) + \varepsilon + \varepsilon(1 - \varepsilon)^n m_n(U)
\]

\[
\leq \sum_{i=1}^{\infty} m_n(h(B_i)) + \varepsilon + \varepsilon(1 - \varepsilon)^n m_n(U)
\]

\[
\leq \sum_{i=1}^{\infty} (1 + \varepsilon)^n |\det Dh(x)| m_n(B_i) + C\varepsilon
\]

\[
= \sum_{i=1}^{\infty} (1 + \varepsilon)^n \int_{B_i} |\det Dh(x)| \, dm_n + C\varepsilon
\]

\[
\leq \sum_{i=1}^{\infty} (1 + \varepsilon)^n \int_{B_i} |\det Dh(x)| \, dm_n + C\varepsilon
\]

\[
\leq \int_V |\det Dh(x)| \, dm_n (1 + \varepsilon)^n + C\varepsilon
\]

\[
\leq \left( \int_A |\det Dh(x)| \, dm_n + \int_{V \setminus A} |\det Dh(x)| \, dm_n \right) (1 + \varepsilon)^n + C\varepsilon
\]
\[
\leq \int_A |\det Dh(x)|\,dm_n (1 + \varepsilon)^n + \varepsilon k (1 + \varepsilon)^n + C\varepsilon
\]

Since \( \varepsilon \) is arbitrary, the ends of this string and \( m_n (h(A)) + \varepsilon + \varepsilon (1 - \varepsilon)^n m_n (V) \) in the middle show that

\[
m_n (h(A)) = \int_A |Dh(x)|\,dm_n
\]

Now we remove the assumption **. Letting \( A \) be arbitrary and measurable, let \( A_k \equiv A \cap \{x : |\det Dh(x)| \leq k\} \)

From what was just shown,

\[
m_n (h(A_k)) = \int_{A_k} |Dh(x)|\,dm_n
\]

Let \( k \rightarrow \infty \) and use monotone convergence theorem to prove the lemma. 

**Corollary 21.5.4** Let \( h : U \rightarrow \mathbb{R}^n \) where \( U \) is an open set and suppose that \( Dh(x) \) exists and is invertible and continuous on \( U \) and that \( h \) is one to one on \( U \). Let \( A \subseteq U \) be a Lebesgue measurable set. Then

\[
m_n (h(A)) = \int_A |Dh(x)|\,dm_n
\]

**Proof:** Let \( A \subseteq U \) and let \( A_k \equiv B (0, k) \cap A, U_k \equiv B (0, k) \cap U \). Then the above lemma shows \( m_n (h(A)) = \int_{A_k} |Dh(x)|\,dm_n \).

Now use the monotone convergence theorem to obtain the conclusion of the lemma. 

**Theorem 21.5.5** Let \( h : U \rightarrow \mathbb{R}^n \) where \( U \) is an open set and suppose that \( Dh(x) \) exists and is continuous and invertible on \( U \) and that \( h \) is one to one on \( U \). Then if \( f \geq 0 \) and is Lebesgue measurable,

\[
\int_{h(U)} f(y)\,dm_n = \int_U f(h(x)) |\det Dh(x)|\,dm_n
\]

**Proof:** Let \( f(y) = \lim_{k \rightarrow \infty} s_k(y) \) where \( s_k \) is a nonnegative simple function. Say

\[
s_k(y) = \sum_{i=1}^{m_k} c_i^k \chi_{F_i^k}(y)
\]

Then

\[
s_k(h(x)) = \sum_{i=1}^{m_k} c_i^k \chi_{h^{-1}(F_i^k)}(x)
\]

It follows from the above corollary that

\[
\int_U s_k(h(x)) |\det Dh(x)|\,dm_n = \sum_{i=1}^{m_k} c_i^k \int_{h^{-1}(F_i^k)} |\det Dh(x)|\,dm_n
\]

\[
= \sum_{i=1}^{m_k} c_i^k m_n (F_i^k) = \int_{h(U)} \sum_{i=1}^{m_k} c_i^k \chi_{F_i^k}(y)\,dm_n = \int_{h(U)} s_k(y)\,dm_n
\]

Now apply monotone convergence theorem to obtain the desired result. 

It is a good idea to remove the requirement that \( \det Dh(x) \neq 0 \). This is also fairly easy from the Besicovitch covering theorem.

The following is Sard's lemma. In the proof, it does not matter which norm you use in defining balls but it may be easiest to consider the norm \( ||x|| \equiv \max \{|x_i|, i = 1, \cdots , n\} \).
Lemma 21.5.6 (Sard) Let $U$ be an open set in $\mathbb{R}^n$ and let $h: U \to \mathbb{R}^n$ be differentiable. Let

$$Z \equiv \{x \in U : \det D h(x) = 0\}.$$

Then $m_n(h(Z)) = 0$.

Proof: For convenience, assume the balls in the following argument come from $\|\cdot\|_\infty$. First note that $Z$ is a Borel set because $h$ is continuous and so the component functions of the Jacobian matrix are each Borel measurable. Hence the determinant is also Borel measurable.

Suppose that $U$ is a bounded open set. Let $\varepsilon > 0$ be given. Also let $V \supseteq Z$ with $V \subseteq U$ open, and

$$m_n(Z) + \varepsilon > m_n(V).$$

Now let $x \in Z$. Then since $h$ is differentiable at $x$, there exists $\delta_x > 0$ such that if $r < \delta_x$, then $B(x, r) \subseteq V$ and also,

$$h(B(x, r)) \subseteq h(x) + D h(x)(B(0, r)) + B(0, r\eta), \quad \eta < 1.$$

Regard $D h(x)$ as an $m \times m$ matrix, the matrix of the linear transformation $D h(x)$ with respect to the usual coordinates. Since $x \in Z$, it follows that there exists an invertible matrix $A$ such that $A D h(x)$ is in row reduced echelon form with a row of zeros on the bottom. Therefore,

$$m_n(A(h(B(x, r)))) \leq m_n(A D h(x)(B(0, r)) + AB(0, r\eta)) \quad (21.19)$$

The diameter of $A D h(x)(B(0, r))$ is no larger than $|A||D h(x)||2r$ and it lies in $\mathbb{R}^{n-1} \times \{0\}$. The diameter of $AB(0, r\eta)$ is no more than $|A|(2r\eta)$. Therefore, the measure of the right side is no more than

$$
\begin{align*}
&\left[|A||D h(x)||2r + |A|(2\eta)\right]^{n-1} r^n
\leq C(|A||D h(x)||2) r^n \eta
\end{align*}
$$

Hence from the change of variables formula for linear maps,

$$m_n(h(B(x, r))) \leq \eta \frac{C(|A||D h(x)||)}{|\det(A)|} m_n(B(x, r))$$

Then letting $\delta_x$ be still smaller if necessary, corresponding to sufficiently small $\eta$,

$$m_n(h(B(x, r))) \leq \varepsilon m_n(B(x, r))$$

The balls of this form constitute a Vitali cover of $Z$. Hence, by the Vitali covering theorem Corollary 21.3.6, there exists $\{B_i\}_{i=1}^\infty, B_i = B_i(x_i, r_i)$, a collection of disjoint balls, each of which is contained in $V$, such that $m_n(h(B_i)) \leq \varepsilon m_n(B_i)$ and $m_n(Z \cup \cup_i B_i) = 0$. Hence from Proposition 21.3.8,

$$m_n(h(Z) \cup \cup_i h(B_i)) \leq m_n(h(Z \cup \cup_i B_i)) = 0$$

Therefore,

$$m_n(h(Z)) \leq \sum_i m_n(h(B_i)) \leq \varepsilon \sum_i m_n(B_i) \leq \varepsilon (m_n(V) \leq \varepsilon (m_n(Z) + \varepsilon).$$

Since $\varepsilon$ is arbitrary, this shows $m_n(h(Z)) = 0$. What if $U$ is not bounded? Then consider $Z_m = Z \cap B(0, m)$. From what was just shown, $h(Z_m)$ has measure 0 and so it follows that $h(Z)$ also does, being the countable union of sets of measure zero.

Now here is a better change of variables theorem.
Theorem 21.5.7 Let $h : U \to \mathbb{R}^n$ where $U$ is an open set and suppose that $Dh(x)$ exists and is continuous on $U$ and that $h$ is one to one on $U$. Then if $f \geq 0$ and is Lebesgue measurable,

$$\int_{h(U)} f(y) \, dm_n = \int_{h(U)} f(h(x)) \, |\det Dh(x)| \, dm_n$$

**Proof:** Let $U_+ = \{x \in U : |\det Dh(x)| > 0\}$, an open set, and $U_0 = \{x \in U : |\det Dh(x)| = 0\}$. Then, using that $h$ is one to one along with Lemma 4.4.8 and Theorem 4.4.9,

$$\int_{h(U)} f(y) \, dm_n = \int_{h(U_+)} f(y) \, dm_n = \int_{U_+} f(h(x)) \, |\det Dh(x)| \, dm_n = \int_{U} f(h(x)) \, |\det Dh(x)| \, dm_n \tag{

Now suppose $h$ is only $C^1$, not necessarily one to one. For

$$U_+ = \{x \in U : |\det Dh(x)| > 0\}$$

and $Z$ the set where $|\det Dh(x)| = 0$, Lemma 4.4.8 implies $m_n(h(Z)) = 0$. For $x \in U_+$, the inverse function theorem implies there exists an open set $B_x \subseteq U_+$, such that $h$ is one to one on $B_x$.

Let $\{B_i\}$ be a countable subset of $\{B_x\}_{x \in U_+}$ such that $U_+ = \bigcup_{i=1}^{\infty} B_i$. Let $E_1 = B_1$. If $E_1, \cdots, E_k$ have been chosen, $E_{k+1} = B_{k+1} \setminus \bigcup_{i=1}^{k} E_i$. Thus

$$\bigcup_{i=1}^{\infty} E_i = U_+, \quad h \text{ is one to one on } E_i, \quad E_i \cap E_j = \emptyset,$$

and each $E_i$ is a Borel set contained in the open set $B_i$. Now define

$$n(y) = \sum_{i=1}^{\infty} X_{h(E_i)}(y) + X_{h(Z)}(y).$$

The set $h(E_i), h(Z)$ are measurable by Proposition 4.5.1. Thus $n(\cdot)$ is measurable.

**Lemma 21.5.8** Let $F \subseteq h(U)$ be measurable. Then

$$\int_{h(U)} n(y) X_F(y) \, dm_n = \int_{U} X_F(h(x)) |\det Dh(x)| \, dm_n.$$

**Proof:** Using Lemma 4.4.8 and the Monotone Convergence Theorem

$$\int_{h(U)} n(y) X_F(y) \, dm_n = \int_{h(U)} \left( \sum_{i=1}^{\infty} X_{h(E_i)}(y) + X_{h(Z)}(y) \right) X_F(y) \, dm_n$$

$$= \sum_{i=1}^{\infty} \int_{h(U)} X_{h(E_i)}(y) X_F(y) \, dm_n$$

$$= \sum_{i=1}^{\infty} \int_{h(B_i)} X_{h(E_i)}(y) X_F(y) \, dm_n$$

$$= \sum_{i=1}^{\infty} \int_{B_i} X_{E_i}(x) X_F(h(x)) \, |\det Dh(x)| \, dm_n$$

$$= \sum_{i=1}^{\infty} \int_{U} X_{E_i}(x) X_F(h(x)) \, |\det Dh(x)| \, dm_n$$

$$= \int_{U} \sum_{i=1}^{\infty} X_{E_i}(x) X_F(h(x)) \, |\det Dh(x)| \, dm_n$$
\[ = \int_{U_+} \chi_F(h(x)) |\det Dh(x)| dm_n = \int_U \chi_F(h(x)) |\det Dh(x)| dm_n. \]

**Definition 21.5.9** For \( y \in h(U) \), define a function, \#, according to the formula

\[ \#(y) \equiv \text{number of elements in } h^{-1}(y). \]

Observe that

\[ \#(y) = n(y) \text{ a.e.} \quad (21.20) \]

because \( n(y) = \#(y) \) if \( y \notin h(Z) \), a set of measure 0. Therefore, \# is a measurable function because of completeness of Lebesgue measure.

**Theorem 21.5.10** Let \( g \geq 0 \), \( g \) measurable, and let \( h \) be \( C^1(U) \). Then

\[ \int_{h(U)} \#(y)g(y) dm_n = \int_U g(h(x)) |\det Dh(x)| dm_n. \quad (21.21) \]

**Proof:** From (21.20) and Lemma 21.5.8 holds for all \( g \), a nonnegative simple function. Approximating an arbitrary measurable nonnegative function, \( g \), with an increasing pointwise convergent sequence of simple functions and using the monotone convergence theorem, yields (21.21) for an arbitrary nonnegative measurable function, \( g \).

### 21.6 Exercises

1. Suppose \( A \subseteq \mathbb{R}^n \) is covered by a finite collection of Balls, \( \mathcal{F} \). Show that then there exists a disjoint collection of these balls, \( \{B_i\}_{i=1}^m \), such that \( A \subseteq \bigcup_{i=1}^m B_i \) where \( B_i \) has the same center as \( B_i \) but 3 times the radius. **Hint:** Since the collection of balls is finite, they can be arranged in order of decreasing radius.

2. In Problem 8, you showed that if \( f \in L^1(\mathbb{R}^n) \), there exists \( h \) which is continuous and equal to 0 off some compact set such that

\[ \int |f - h| dm < \varepsilon \]

Define \( f_y(x) \equiv f(x - y) \). Explain why \( f_y \) is Lebesgue measurable and

\[ \int |f_y| dm_n = \int |f| dm_n \]

Now justify the following formula.

\[ \int |f_y - f| dm_n \leq \int |f_y - h_y| dm_n + \int |h_y - h| dm_n + \int |h - f| dm_n \]

\[ \leq 2\varepsilon + \int |h_y - h| dm_n \]

Now explain why the last term is less than \( \varepsilon \) if \( \|y\| \) is small enough. Explain continuity of translation in \( L^1(\mathbb{R}^n) \) which says that

\[ \lim_{y \to 0} \int_{\mathbb{R}^n} |f_y - f| dm_n = 0 \]

3. This problem will help to understand that a certain kind of function exists.

\[ f(x) = \begin{cases} 
  e^{-1/x^2} & \text{if } x \neq 0 \\
  0 & \text{if } x = 0
\end{cases} \]

show that \( f \) is infinitely differentiable. Note that you only need to be concerned with what happens at 0. There is no question elsewhere. This is a little fussy but is not too hard.
4. ↑Let \( f(x) \) be as given above. Now let

\[
\tilde{f}(x) = \begin{cases} 
  f(x) & \text{if } x \leq 0 \\
  0 & \text{if } x > 0
\end{cases}
\]

Show that \( \tilde{f}(x) \) is also infinitely differentiable. Now consider \( r > 0 \) and define \( g(x) = \tilde{f}(-(x-r)) \tilde{f}(x+r) \). show that \( g \) is infinitely differentiable and vanishes for \( |x| \geq r \). Let \( \psi(x) = \prod_{k=1}^{n} g(x_k) \). For \( U = B(0, 2r) \) with the norm given by \( ||x|| = \max\{ |x_k|, k \leq n \} \), show that \( \psi \in C_c^\infty(U) \).

5. ↑Using the above problem, let \( \psi \in C_c^\infty(B(0, 1)) \). Also let \( \psi \geq 0 \) as in the above problem. Show there exists \( \psi \geq 0 \) such that \( \psi \in C_c^\infty(B(0, 1)) \) and \( \int \psi dm_n = 1 \). Now define

\[
\psi_k(x) \equiv k^n \psi(kx)
\]

Show that \( \psi_k \) equals zero off a compact subset of \( B(0, \frac{1}{k}) \) and \( \int \psi_k dm_n = 1 \). We say that \( \text{spt}(\psi_k) \subseteq B(0, \frac{1}{k}) \). \( \text{spt}(f) \) is defined as the closure of the set on which \( f \) is not equal to zero. Such a sequence of functions as just defined \( \{\psi_k\} \) where \( \int \psi_k dm_n = 1 \) and \( \psi_k \geq 0 \) and \( \text{spt}(\psi_k) \subseteq B(0, \frac{1}{k}) \) is called a mollifier.

6. ↑It is important to be able to approximate functions with those which are infinitely differentiable. Suppose \( f \in L^1(\mathbb{R}^n) \) and let \( \{\psi_k\} \) be a mollifier as above. We define the convolution as follows.

\[
f*\psi_k(x) \equiv \int f(x-y) \psi_k(y) dm_n(y)
\]

Here the notation means that the variable of integration is \( y \). Show that \( f*\psi_k(x) \) exists and equals

\[
\int \psi_k(x-y) f(y) dm_n(y)
\]

Now show using the dominated convergence theorem that \( f*\psi_k \) is infinitely differentiable. Next show that

\[
\lim_{k \to \infty} \int |f(x) - f*\psi_k(x)| dm_n = 0
\]

Thus, in terms of being close in \( L^1(\mathbb{R}^n) \), every function in \( L^1(\mathbb{R}^n) \) is close to one which is infinitely differentiable.

7. ↑From Problem 6 above and \( f \in L^1(\mathbb{R}^n) \), there exists \( h \in C_c(\mathbb{R}^n) \), continuous and \( \text{spt}(h) \) a compact set, such that

\[
\int |f - h| dm_n < \varepsilon
\]

Now consider \( h*\psi_k \). Show that this function is in \( C_c^\infty(\text{spt}(h) + B(0, \frac{1}{k})) \). The notation means you start with the compact set \( \text{spt}(h) \) and fatten it up by adding the set \( B(0,\frac{1}{k}) \).

It means \( x+y \) such that \( x \in \text{spt}(h) \) and \( y \in B(0,\frac{1}{k}) \). Show the following. For all \( k \) large enough,

\[
\int |f - h*\psi_k| dm_n < \varepsilon
\]

so one can approximate with a function which is infinitely differentiable and also has compact support. Also show that \( h*\psi_k \) converges uniformly to \( h \). If \( h \) is a function in \( C^k(\mathbb{R}^k) \) in addition to being continuous with compact support, show that for each \( |\alpha| \leq k \), \( D^\alpha (h*\psi_k) \to D^\alpha h \) uniformly. **Hint:** If you do this for a single partial derivative, you will see how it works in general.
8. Let \( f \in L^1(\mathbb{R}) \). Show that
\[
\lim_{k \to \infty} \int f(x) \sin(kx) \, dm = 0
\]
**Hint:** Use the result of the above problem to obtain \( g \in C_c^\infty(\mathbb{R}) \), continuous and zero off a compact set, such that
\[
\int |f - g| \, dm < \varepsilon
\]
Then show that
\[
\lim_{k \to \infty} \int g(x) \sin(kx) \, dm(x) = 0
\]
You can do this by integration by parts. Then consider this.
\[
\left| \int f(x) \sin(kx) \, dm \right| = \left| \int f(x) \sin(kx) \, dm - \int g(x) \sin(kx) \, dm \right|
+ \left| \int g(x) \sin(kx) \, dm \right|
\leq \int |f - g| \, dm + \left| \int g(x) \sin(kx) \, dm \right|
\]
This is the celebrated Riemann Lebesgue lemma which is the basis for all theorems about pointwise convergence of Fourier series.

9. As another application of theory of regularity, here is a very important result. Suppose \( f \in L^1(\mathbb{R}^n) \) and for every \( \phi \in C_c^\infty(\mathbb{R}^n) \)
\[
\int f \phi \, dm = 0
\]
Show that then it follows that \( f(x) = 0 \) for a.e. \( x \). That is, there is a set of measure zero such that off this set \( f \) equals 0. **Hint:** What you can do is to let \( E \) be a measurable which is bounded and let \( K_k \subseteq E \subseteq V_k \) where \( m_k(V_k \setminus K_k) < 2^{-k} \). Here \( K_k \) is compact and \( V_k \) is open. By an earlier exercise, Problem 12 on Page 474, there exists a function \( \phi_k \) which is continuous, has values in \([0,1]\) equals 1 on \( K_k \) and \( \text{spt}(\phi_k) \subseteq V_k \). To get this last part, show there exists \( W_k \) open such that \( W_k \subseteq V_k \) and \( W_k \) contains \( K_k \). Then you use the problem to get \( \text{spt}(\phi_k) \subseteq \overline{W_k} \). Now you form \( \eta_k = \phi_k * \psi_l \) where \( \{\psi_l\} \) is a mollifier. Show that for \( l \) large enough, \( \eta_k \) has values in \([0,1]\), \( \text{spt}(\eta_k) \subseteq V_k \) and \( \eta_k \in C_c^\infty(V_k) \). Now explain why \( \eta_k \to \chi_E \) off a set of measure zero. Then
\[
\left| \int f \chi_E \, dm \right| = \left| \int f(\chi_E - \eta_k) \, dm \right| + \left| \int f \eta_k \, dm \right|
\leq \int |f - g| \, dm + \int g(x) \sin(kx) \, dm
\]
Now explain why this converges to 0 on the right. This will involve the dominated convergence theorem. Conclude that \( \int f \chi_E \, dm = 0 \) for every bounded measurable set \( E \). Show that this implies that \( \int f \chi_E \, dm = 0 \) for every measurable \( E \). Explain why this requires \( f = 0 \) a.e. The result which gets used over and over in all of this is the dominated convergence theorem.

10. This is from the advanced calculus book by Apostol. Justify the following argument using convergence theorems.
\[
F(x) \equiv \int_0^1 \frac{e^{-x^2(1+t^2)}}{1+t^2} \, dt + \left( \int_0^x e^{-t^2} \, dt \right)^2
\]
Then you can take the derivative $\frac{d}{dx}$ and obtain

$$F' (x) = \int_0^1 e^{-x^2(1+t^2)} (-2x) \left( 1 + \frac{t^2}{1+t^2} \right) dt + 2 \left( \int_0^x e^{-x^2} dt \right) e^{-x^2}$$

$$= \int_0^1 e^{-x^2(1+t^2)} (-2x) dt + 2 \left( \int_0^x e^{-t^2} dt \right) e^{-x^2}$$

Why can you take the derivative on the inside of the integral? Change the variable in the second integral, letting $t = sx$. Thus $dt = xds$. Then the above reduces to

$$= \int_0^1 e^{-x^2(1+t^2)} (-2x) dt + 2x \left( \int_0^1 e^{-x^2(x^2+s^2)} ds \right) e^{-x^2}$$

$$= (-2x) \int_0^1 e^{-x^2(1+t^2)} dt + 2x \left( \int_0^1 e^{-(x^2+1)x^2} ds \right) = 0$$

and so this function of $x$ is constant. However,

$$F (0) = \int_0^1 \frac{dt}{1+t^2} = \frac{1}{4} \pi$$

Now you let $x \to \infty$. What happens to that first integral? It equals

$$e^{-x^2} \int_0^1 e^{-x^2} \frac{t^2}{1+t^2} dt \leq e^{-x^2} \left( \int_0^1 \frac{1}{1+t^2} dt \right)$$

and so it obviously converges to 0 as $x \to \infty$. Therefore, taking a limit yields

$$\left( \int_0^\infty e^{-t^2} dt \right)^2 = \frac{\pi}{4}, \int_0^\infty e^{-t^2} dt = \frac{\sqrt{\pi}}{2}$$

11. The Dini derivates are as follows. In these formulas, $f$ is a real valued function defined on $\mathbb{R}$.

$$D^+ f (x) \equiv \lim_{h \to 0^+} \sup_{h < 0} \frac{f (x + h) - f (x)}{h}, D^+ f (x) \equiv \lim_{h \to 0^+} \inf_{h < 0} \frac{f (x + h) - f (x)}{h}$$

$$D^- f (x) \equiv \lim_{h \to 0^+} \sup_{h > 0} \frac{f (x) - f (x - h)}{h}, D^- f (x) \equiv \lim_{h \to 0^+} \inf_{h > 0} \frac{f (x) - f (x - h)}{h}$$

Thus when these are all equal, the function has a derivative. Now suppose $f$ is an increasing function. Let

$$N_{ab} = \{ x : D^+ f (x) > b > a > D^- f (x) \}, a \geq 0$$

Let $V$ be an open set which contains $N_{ab} \cap (-r, r) \equiv N_{ab}^r$ such that

$$m (V \setminus (N_{ab} \cap (-r, r))) < \varepsilon$$

Then explain why there exist disjoint intervals $[a_i, b_i]$ such that

$$m (N_{ab}^r \setminus \cup_i [a_i, b_i]) = m (N_{ab}^r \setminus \cup_i (a_i, b_i)) = 0$$

and

$$f (b_i) - f (a_i) \leq am (a_i, b_i)$$

each interval being contained in $V \cap (-r, r)$. Thus you have

$$m (N_{ab}^r) = m (\cup_i N_{ab}^r \cap (a_i, b_i)).$$
Next show there exist disjoint intervals \((a_j, b_j)\) such that each of these is contained in some \((a_i, b_i)\), the \((a_j, b_j)\) are disjoint, \(f(b_j) - f(a_j) \geq bm(a_j, b_j)\), and \(\sum_j m(N_{ab}^r \cap (a_j, b_j)) = m(N_{ab}^r)\). Then you have the following thanks to the fact that \(f\) is increasing.

\[
a(m(N_{ab}^r) + \varepsilon) > am(V) \geq a \sum_i (b_i - a_i) > \sum_i f(b_i) - f(a_i) \\
\geq \sum_j f(b_j) - f(a_j) \geq b \sum_j b_j - a_j \\
\geq b \sum_j m(N_{ab}^r \cap (a_j, b_j)) = bm(N_{ab}^r)
\]

and since \(\varepsilon > 0\),

\[
am(N_{ab}^r) \geq bm(N_{ab}^r)
\]

showing that \(m(N_{ab}^r) = 0\). This is for any \(r\) and so \(m(N_{ab}) = 0\). Thus the derivative from the right exists for a.e. \(x\) by taking the complement of the union of the \(N_{ab}\) for \(a, b\) nonnegative rational numbers. Now do the same thing to show that the derivative from the left exists a.e. and finally, show that \(D_-f(x) = D^+f(x)\) for almost a.e. \(x\). Off the union of these three exceptional sets of measure zero all the derivatives are the same and so the derivative of \(f\) exists a.e. In other words, an increasing function has a derivative a.e.

12. This problem is on Eggoroff’s theorem. Suppose you have a measure space \((\Omega, \mathcal{F}, \mu)\) where \(\mu(\Omega) < \infty\). Also suppose that \(\{f_k\}\) is a sequence of measurable, complex valued functions which converge to \(f\) pointwise. Then Eggoroff’s theorem says that for any \(\varepsilon > 0\) there is a set \(N\) with \(\mu(N) < \varepsilon\) and convergence is uniform on \(N^C\).

(a) Define \(E_{mk} \equiv \bigcup_{r=m}^{\infty} \{\omega: |f(\omega) - f_r(\omega)| > \frac{1}{k}\}\). Show that \(E_{mk} \supseteq E_{(m+1)k}\) for all \(m\) and that \(\bigcap_m E_{mk} = \emptyset\).

(b) Show that there exists \(m(k)\) such that \(\mu(E_{m(k)k}) < \varepsilon 2^{-k}\).

(c) Let \(N \equiv \bigcup_{k=1}^{\infty} E_{m(k)k}\). Explain why \(\mu(N) < \varepsilon\) and that for all \(\omega \notin N^C\), if \(r > m(k)\), then \(|f(\omega) - f_r(\omega)| \leq \frac{1}{k}\). Thus uniform convergence takes place on \(N^C\).

13. Suppose you have a sequence \(\{f_n\}\) which converges uniformly on each of sets \(A_1, \cdots, A_n\). Why does the sequence converge uniformly on \(\bigcup_{i=1}^{n} A_i\)?

14. Now suppose you have a \((\Omega, \mathcal{F}, \mu)\) where \(\mu\) is a finite Radon measure and \(\Omega\) is a metric space. For example, you could have Lebesgue measure for \(\Omega\) a bounded open subset of \(\mathbb{R}^n\). Suppose you have \(f\) has nonnegative real values for all \(\omega\) and is measurable. Then Lusin’s theorem says that for every \(\varepsilon > 0\), there exists an open set \(V\) with measure less than \(\varepsilon\) and a continuous function defined on \(\Omega\) such that \(f(\omega) = g(\omega)\) for all \(\omega \notin V\).

(a) By Lemma 24.12.1, there exists an increasing sequence \(\{f_n\} \subseteq C_c(\Omega)\) which converges to \(f\) off a set \(N\) of measure zero. Use Eggoroff’s theorem to enlarge \(N\) to \(\tilde{N}\) such that \(\mu(\tilde{N}) < \frac{\varepsilon}{2}\) and convergence is uniform off \(\tilde{N}\).

(b) Next use outer regularity to obtain open \(V \supseteq \tilde{N}\) having measure less than \(\varepsilon\). Thus \(\{f_n\}\) converges uniformly on \(V^C\). Therefore, that which it converges to is continuous on \(V^C\) a closed set. Now use the Tietze extension theorem.

15. Let \(h: U \to \mathbb{R}^n\) be differentiable for \(U\) an open set in \(\mathbb{R}^n\). Thus, as explained above, \(h\) takes measurable sets to measurable sets. Suppose \(h\) is one to one. For \(E\) a measurable subset of \(U\) define

\[
\mu(E) \equiv m_n(h(E))
\]

Show that \(\mu\) is a measure and whenever \(m_n(E) = 0\), it follows that \(\mu(E) = 0\). When this happens, we say that \(\mu \ll m_n\). Great things can be said in this situation. It is this which allows one to dispense with the assumption in the change of variables formula that \(h\) is \(C^1\).
16. Say you have a change of variables formula which says that

$$\int_{h(U)} f(y) \, dm_n = \int_U f(h(x)) |\det Dh(x)| \, dm_n$$

and that this holds for all $f \geq 0$ and Lebesgue measurable. Show that the formula continues to hold if $f \in L^1(h(U))$. **Hint:** Apply what is known to the positive and negative parts of the real and imaginary parts.
Chapter 22

The $L^p$ Spaces

This book is about linear algebra and things which are related to this subject. Nearly all of it is on finite dimensional vector spaces. Now it is time to include some of the most useful function spaces. They are infinite dimensional vector spaces.

22.1 Basic Inequalities And Properties

One of the main applications of the Lebesgue integral is to the study of various sorts of functions space. These are vector spaces whose elements are functions of various types. One of the most important examples of a function space is the space of measurable functions whose absolute values are $p^{th}$ power integrable where $p \geq 1$. These spaces, referred to as $L^p$ spaces, are very useful in applications. In the chapter $(\Omega, S, \mu)$ will be a measure space.

Definition 22.1.1 Let $1 \leq p < \infty$. Define

$$L^p(\Omega) \equiv \{ f : f \text{ is measurable and } \int_\Omega |f(\omega)|^p d\mu < \infty \}$$

For each $p > 1$ define $q$ by

$$\frac{1}{p} + \frac{1}{q} = 1.$$ 

Often one uses $p'$ instead of $q$ in this context.

$L^p(\Omega)$ is a vector space and has a norm. This is similar to the situation for $\mathbb{R}^n$ but the proof requires the following fundamental inequality.

Theorem 22.1.2 (Holder’s inequality) If $f$ and $g$ are measurable functions, then if $p > 1$,

$$\int |f| |g| \, d\mu \leq \left( \int |f|^p \, d\mu \right)^{\frac{1}{p}} \left( \int |g|^q \, d\mu \right)^{\frac{1}{q}}.$$ \hspace{1cm} (22.1)

Proof: First here is a proof of Young’s inequality.

Lemma 22.1.3 If $p > 1$, and $0 \leq a, b$ then $ab \leq \frac{a^n}{p} + \frac{b^n}{q}$. 

535
**Proof:** Consider the following picture:

From this picture, the sum of the area between the $x$ axis and the curve added to the area between the $t$ axis and the curve is at least as large as $ab$. Using beginning calculus, this is equivalent to the following inequality.

$$ab \leq \int_0^a t^{p-1} dt + \int_0^b x^{q-1} dx = \frac{a^p}{p} + \frac{b^q}{q}.$$  

The above picture represents the situation which occurs when $p > 2$ because the graph of the function is concave up. If $2 > p > 1$ the graph would be concave down or a straight line. You should verify that the same argument holds in these cases just as well. In fact, the only thing which matters in the above inequality is that the function $x = t^{p-1}$ be strictly increasing.

Note equality occurs when $a^p = b^q$.

Here is an alternate proof.

**Lemma 22.1.4** For $a, b \geq 0$,

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

and equality occurs when if and only if $a^p = b^q$.

**Proof:** If $b = 0$, the inequality is obvious. Fix $b > 0$ and consider

$$f(a) = \frac{a^p}{p} + \frac{b^q}{q} - ab.$$  

Then $f'(a) = a^{p-1} - b.$ This is negative when $a < b^{1/(p-1)}$ and is positive when $a > b^{1/(p-1)}$. Therefore, $f$ has a minimum when $a = b^{1/(p-1)}$. In other words, when $a^p = b^{p/(p-1)} = b^q$ since $1/p + 1/q = 1$. Thus the minimum value of $f$ is

$$\frac{b^q}{p} + \frac{b^q}{q} - b^{1/(p-1)}b = b^q - b^q = 0.$$  

It follows $f \geq 0$ and this yields the desired inequality.

**Proof of Holder’s inequality:** If either $\int |f|^p d\mu$ or $\int |g|^p d\mu$ equals $\infty$, the inequality (22.1) is obviously valid because $\infty \geq$ anything. If either $\int |f|^p d\mu$ or $\int |g|^p d\mu$ equals 0, then $f = 0$ a.e. or that $g = 0$ a.e. and so in this case the left side of the inequality equals 0 and so the inequality is therefore true. Therefore assume both $\int |f|^p d\mu$ and $\int |g|^p d\mu$ are less than $\infty$ and not equal to 0. Let

$$\left(\frac{\int |f|^p d\mu}{I(f)}\right)^{1/p} = I(f)$$

and let $\left(\frac{\int |g|^p d\mu}{I(g)}\right)^{1/q} = I(g)$. Then using the lemma,

$$\int \frac{|f|}{I(f)} \frac{|g|}{I(g)} d\mu \leq \frac{1}{p} \int \frac{|f|^p}{I(f)^p} d\mu + \frac{1}{q} \int \frac{|g|^q}{I(g)^q} d\mu = 1.$$
Hence,
\[
\int |f| |g| d\mu \leq I(f) I(g) = \left( \int |f|^p d\mu \right)^{1/p} \left( \int |g|^q d\mu \right)^{1/q}.
\]
This proves Holder’s inequality.

The following lemma will be needed.

**Lemma 22.1.5** Suppose \( x, y \in \mathbb{C} \). Then
\[
|x + y|^p \leq 2^{p-1} (|x|^p + |y|^p).
\]

**Proof:** The function \( f(t) = t^p \) is concave up for \( t \geq 0 \) because \( p > 1 \). Therefore, the secant line joining two points on the graph of this function must lie above the graph of the function. This is illustrated in the following picture.

Now as shown above,
\[
\left( \frac{|x| + |y|}{2} \right)^p \leq \frac{|x|^p + |y|^p}{2}
\]
which implies
\[
|x + y|^p \leq (|x| + |y|)^p \leq 2^{p-1} (|x|^p + |y|^p)
\]
and this proves the lemma.

Note that if \( y = \phi(x) \) is any function for which the graph of \( \phi \) is concave up, you could get a similar inequality by the same argument.

**Corollary 22.1.6** *(Minkowski inequality)* Let \( 1 \leq p < \infty \). Then
\[
\left( \int |f + g|^p d\mu \right)^{1/p} \leq \left( \int |f|^p d\mu \right)^{1/p} + \left( \int |g|^p d\mu \right)^{1/p}.
\]

**Proof:** If \( p = 1 \), this is obvious because it is just the triangle inequality. Let \( p > 1 \). Without loss of generality, assume
\[
\left( \int |f|^p d\mu \right)^{1/p} + \left( \int |g|^p d\mu \right)^{1/p} < \infty
\]
and \( \left( \int |f + g|^p d\mu \right)^{1/p} \neq 0 \) or there is nothing to prove. Therefore, using the above lemma,
\[
\int |f + g|^p d\mu \leq 2^{p-1} \left( \int |f|^p + |g|^p d\mu \right) < \infty.
\]
Now \( |f(\omega) + g(\omega)|^p \leq |f(\omega) + g(\omega)|^{p-1} (|f(\omega)| + |g(\omega)|) \). Also, it follows from the definition of \( p \) and \( q \) that \( p - 1 = \frac{p}{q} \). Therefore, using this and Holder’s inequality,
\[
\int |f + g|^p d\mu \leq
\]
\[
\int |f + g|^{p-1}|f|d\mu + \int |f + g|^{p-1}|g|d\mu = \int |f + g|^\frac{p}{2}|f|d\mu + \int |f + g|^\frac{p}{2}|g|d\mu \leq \left( \int |f + g|^p d\mu \right)^\frac{1}{p} + \left( \int |f|^p d\mu \right)^\frac{1}{p} + \left( \int |g|^p d\mu \right)^\frac{1}{p}.
\]

Dividing both sides by \( \left( \int |f + g|^p d\mu \right)^\frac{1}{p} \) yields \( \frac{1}{p} \) This proves the corollary.

The following follows immediately from the above.

**Corollary 22.1.7** Let \( f_i \in L^p(\Omega) \) for \( i = 1, 2, \ldots, n \). Then

\[
\left( \int \left| \sum_{i=1}^{n} f_i \right|^p d\mu \right)^{1/p} \leq \sum_{i=1}^{n} \left( \int |f_i|^p \right)^{1/p}.
\]

This shows that if \( f, g \in L^p \), then \( f + g \in L^p \). Also, it is clear that if \( a \) is a constant and \( f \in L^p \), then \( af \in L^p \) because

\[
\int |af|^p d\mu = |a|^p \int |f|^p d\mu < \infty.
\]

Thus \( L^p \) is a vector space and

a.) \( \left( \int |f|^p d\mu \right)^{1/p} \geq 0, \left( \int |f|^p d\mu \right)^{1/p} = 0 \) if and only if \( f = 0 \) a.e.

b.) \( \left( \int |a|^p |f|^p d\mu \right)^{1/p} = |a| \left( \int |f|^p d\mu \right)^{1/p} \) if \( a \) is a scalar.

c.) \( \left( \int |f + g|^p d\mu \right)^{1/p} \leq \left( \int |f|^p d\mu \right)^{1/p} + \left( \int |g|^p d\mu \right)^{1/p} \).

f \to \left( \int |f|^p d\mu \right)^{1/p} \) would define a norm if \( \left( \int |f|^p d\mu \right)^{1/p} = 0 \) implied \( f = 0 \). Unfortunately, this is not so because if \( f = 0 \) a.e. but is nonzero on a set of measure zero, \( \left( \int |f|^p d\mu \right)^{1/p} = 0 \) and this is not allowed. However, all the other properties of a norm are available and so a little thing like a set of measure zero will not prevent the consideration of \( L^p \) as a normed vector space if two functions in \( L^p \) which differ only on a set of measure zero are considered the same. That is, an element of \( L^p \) is really an equivalence class of functions where two functions are equivalent if they are equal a.e.

With this convention, here is a definition.

**Definition 22.1.8** Let \( f \in L^p(\Omega) \). Define

\[ ||f||_p \equiv ||f||_{L^p} \equiv \left( \int |f|^p d\mu \right)^{1/p}. \]

Then with this definition and using the convention that elements in \( L^p \) are considered to be the same if they differ only on a set of measure zero, \( || \cdot ||_p \) is a norm on \( L^p(\Omega) \) because if \( ||f||_p = 0 \) then \( f = 0 \) a.e. and so \( f \) is considered to be the zero function because it differs from 0 only on a set of measure zero.

The following is an important definition.

**Definition 22.1.9** A complete normed linear space is called a Banach space.

These spaces are named after Stefan Banach, 1892-1945. Banach spaces are the basic item of study in the subject of functional analysis and will be considered later in this book.

There is a recent biography of Banach, R. Katuža, *The Life of Stefan Banach*, (A. Kostant and W. Woyczyński, translators and editors) Birkhauser, Boston (1996). More information on Banach can also be found in a recent short article written by Douglas Henderson who is in the department of chemistry and biochemistry at BYU.

Banach was born in Austria, worked in Poland and died in the Ukraine but never moved. This is because borders kept changing. There is a rumor that he died in a German concentration camp which is apparently not true. It seems he died after the war of lung cancer.

He was an interesting character. He hated taking examinations so much that he did not receive his undergraduate university degree. Nevertheless, he did become a professor of mathematics due to his important research. He and some friends would meet in a cafe called the Scottish cafe where they wrote on the marble table tops until Banach’s wife supplied them with a notebook which became the ”Scottish notebook” and was eventually published.
Theorem 22.1.10 The following hold for $L^p(\Omega)$

a.) $L^p(\Omega)$ is complete.

b.) If $\{f_n\}$ is a Cauchy sequence in $L^p(\Omega)$, then there exists $f \in L^p(\Omega)$ and a subsequence which converges a.e. to $f \in L^p(\Omega)$, and $\|f_n - f\|_p \to 0$.

Proof: Let $\{f_n\}$ be a Cauchy sequence in $L^p(\Omega)$. This means that for every $\varepsilon > 0$ there exists $N$ such that if $n, m \geq N$, then $\|f_n - f_m\|_p < \varepsilon$. Now select a subsequence as follows. Let $n_1$ be such that $\|f_n - f_{m}\|_p < 2^{-1}$ whenever $n, m \geq n_1$. Let $n_2$ be such that $n_2 > n_1$ and $\|f_n - f_{m}\|_p < 2^{-2}$ whenever $n, m \geq n_2$. If $n_1, \ldots, n_k$ have been chosen, let $n_{k+1} > n_k$ and whenever $n, m \geq n_{k+1}, \|f_n - f_m\|_p < 2^{-(k+1)}$. The subsequence just mentioned is $\{f_{n_k}\}$. Thus, $\|f_{n_k} - f_{n_{k+1}}\|_p < 2^{-k}$. Let $g_{k+1} = f_{n_{k+1}} - f_{n_k}$.

Then by the corollary to Minkowski’s inequality,

$$\infty > \sum_{k=1}^{\infty} \|g_{k+1}\|_p \geq \sum_{k=1}^{m} \|g_{k+1}\|_p \geq \left\| \sum_{k=1}^{m} [g_{k+1}] \right\|_p$$

for all $m$. It follows that

$$\int \left( \sum_{k=1}^{m} |g_{k+1}| \right)^p d\mu \leq \left( \sum_{k=1}^{\infty} \|g_{k+1}\|_p \right)^p < \infty$$

(22.3)

for all $m$ and so the monotone convergence theorem implies that the sum up to $m$ in (22.3) can be replaced by a sum up to $\infty$. Thus,

$$\int \left( \sum_{k=1}^{\infty} |g_{k+1}| \right)^p d\mu < \infty$$

which requires

$$\sum_{k=1}^{\infty} |g_{k+1}(x)| < \infty \text{ a.e. } x.$$

Therefore, $\sum_{k=1}^{\infty} g_{k+1}(x)$ converges for a.e. $x$ because the functions have values in a complete space, $\mathbb{C}$, and this shows the partial sums form a Cauchy sequence. Now let $x$ be such that this sum is finite. Then define

$$f(x) \equiv f_{n_1}(x) + \sum_{k=1}^{\infty} g_{k+1}(x) = \lim_{m \to \infty} f_{n_m}(x)$$

since $\sum_{k=1}^{m} g_{k+1}(x) = f_{n_{m+1}}(x) - f_{n_1}(x)$. Therefore there exists a set, $E$ having measure zero such that

$$\lim_{k \to \infty} f_{n_k}(x) = f(x)$$

for all $x \notin E$. Redefine $f_{n_k}$ to equal 0 on $E$ and let $f(x) = 0$ for $x \in E$. It then follows that $\lim_{k \to \infty} f_{n_k}(x) = f(x)$ for all $x$. By Fatou’s lemma, and the Minkowski inequality,

$$\|f - f_{n_k}\|_p = \left( \int |f - f_{n_k}|^p d\mu \right)^{1/p} \leq \lim_{m \to \infty} \left( \int |f_{n_m} - f_{n_k}|^p d\mu \right)^{1/p} = \lim_{m \to \infty} \inf \|f_{n_m} - f_{n_k}\|_p \leq \lim_{m \to \infty} \sum_{j=k}^{m-1} \|f_{n_{j+1}} - f_{n_j}\|_p \leq \sum_{i=k}^{\infty} \|f_{n_{i+1}} - f_{n_i}\|_p \leq 2^{-(k-1)}.$$  

(22.4)
Therefore, \( f \in L^p(\Omega) \) because
\[
||f||_p \leq ||f - f_{n_k}||_p + ||f_{n_k}||_p < \infty,
\]
and \( \lim_{k \to \infty} ||f_{n_k} - f||_p = 0 \). This proves b).

This has shown \( f_{n_k} \) converges to \( f \) in \( L^p(\Omega) \). It follows the original Cauchy sequence also converges to \( f \) in \( L^p(\Omega) \). This is a general fact that if a subsequence of a Cauchy sequence converges, then so does the original Cauchy sequence. You should give a proof of this.

In working with the \( L^p \) spaces, the following inequality also known as Minkowski’s inequality is very useful. It is similar to the Minkowski inequality for sums. To see this, replace the integral, \( \int_X \) with a finite summation sign and you will see the usual Minkowski inequality or rather the version of it given in Corollary 22.1.7.

To prove this theorem first consider a special case of it in which technical considerations which shed no light on the proof are excluded. For the notion of product measure used here see Problem 6 on Page 504.

**Lemma 22.1.11** Let \((X, \mathcal{S}, \mu)\) and \((Y, \mathcal{F}, \lambda)\) be finite complete measure spaces and let \( f \) be \( \mu \times \lambda \) measurable and uniformly bounded. Then the following inequality is valid for \( p \geq 1 \).

\[
\int_X \left( \int_Y |f(x, y)|^p \, d\lambda \right)^{\frac{1}{p}} \, d\mu \geq \left( \int_Y \left( \int_X |f(x, y)|^p \, d\mu \right)^{\frac{1}{p}} \, d\lambda \right)^{\frac{1}{p}}. \tag{22.5}
\]

**Proof:** Since \( f \) is bounded and \( \mu(X), \lambda(Y) < \infty \),
\[
\left( \int_Y \left( \int_X |f(x, y)|^p \, d\mu \right)^{\frac{1}{p}} \, d\lambda \right)^{\frac{1}{p}} < \infty.
\]

Let
\[
J(y) = \int_X |f(x, y)| \, d\mu.
\]

Note there is no problem in writing this for a.e. \( y \) because \( f \) is product measurable. Then by Fubini’s theorem,
\[
\int_Y \left( \int_X |f(x, y)|^p \, d\mu \right)^{\frac{1}{p}} \, d\lambda = \int_Y J(y)^{p-1} \int_X |f(x, y)| \, d\mu \, d\lambda = \int_X \int_Y J(y)^{p-1} |f(x, y)| \, d\lambda \, d\mu.
\]

Now apply Holder’s inequality in the last integral above and recall \( p - 1 = \frac{p}{q} \). This yields
\[
\int_Y \left( \int_X |f(x, y)| \, d\mu \right)^p \, d\lambda \leq \int_X \left( \int_Y J(y)^p \, d\lambda \right)^{\frac{1}{p}} \left( \int_Y |f(x, y)|^p \, d\lambda \right)^{\frac{1}{p}} \, d\mu = \left( \int_Y J(y)^p \, d\lambda \right)^{\frac{1}{p}} \int_X \left( \int_Y |f(x, y)|^p \, d\lambda \right)^{\frac{1}{p}} \, d\mu = \left( \int_Y \left( \int_X |f(x, y)|^p \, d\mu \right)^{\frac{1}{p}} \, d\lambda \right)^{\frac{1}{p}} \int_X \left( \int_Y |f(x, y)|^p \, d\lambda \right)^{\frac{1}{p}} \, d\mu. \tag{22.6}
\]

Therefore, dividing both sides by the first factor in the above expression,
\[
\left( \int_Y \left( \int_X |f(x, y)|^p \, d\mu \right)^{\frac{1}{p}} \, d\lambda \right)^{\frac{1}{p}} \leq \int_X \left( \int_Y |f(x, y)|^p \, d\lambda \right)^{\frac{1}{p}} \, d\mu. \tag{22.7}
\]

Note that 22.7 holds even if the first factor of 22.5 equals zero. This proves the lemma.

Now consider the case where \( f \) is not assumed to be bounded and where the measure spaces are \( \sigma \) finite.
22.2. Density Considerations

**Theorem 22.1.12** Let \((X,\mathcal{S},\mu)\) and \((Y,\mathcal{F},\lambda)\) be \(\sigma\)-finite measure spaces and let \(f\) be product measurable. Then the following inequality is valid for \(p \geq 1\).

\[
\left( \int_X \left( \int_Y |f(x,y)|^p \, d\lambda \right)^{\frac{1}{p}} \, d\mu \right)^\frac{1}{p} \geq \left( \int_X \left( \int_Y |f(x,y)|^{p} \, d\mu \right)^{\frac{1}{p}} \, d\lambda \right)^\frac{1}{p}.
\] (22.8)

**Proof:** Since the two measure spaces are \(\sigma\)-finite, this means there exist measurable sets, \(X_m\) and \(Y_k\) such that \(X_m \subseteq X_{m+1}\) for all \(m\), \(Y_k \subseteq Y_{k+1}\) for all \(k\), and \(\mu(X_m), \lambda(Y_k) < \infty\). Now define

\[
f_n(x,y) \equiv \begin{cases} f(x,y) & \text{if } |f(x,y)| \leq n \\ n & \text{if } |f(x,y)| > n. \end{cases}
\]

Thus \(f_n\) is uniformly bounded and product measurable. By the above lemma,

\[
\left( \int_X \left( \int_Y |f_n(x,y)|^p \, d\lambda \right)^{\frac{1}{p}} \, d\mu \right)^\frac{1}{p} \geq \left( \int_Y \left( \int_X |f_n(x,y)|^{p} \, d\mu \right)^{\frac{1}{p}} \, d\lambda \right)^\frac{1}{p}.
\] (22.9)

Now observe that \(|f_n(x,y)|\) increases in \(n\) and the pointwise limit is \(|f(x,y)|\). Therefore, using the monotone convergence theorem in \(\mathbb{R}^2\) yields the same inequality with \(f\) replacing \(f_n\). Next let \(k \to \infty\) and use the monotone convergence theorem again to replace \(Y_k\) with \(Y\). Finally let \(m \to \infty\) in what is left to obtain (22.8).

Note that the proof of this theorem depends on two manipulations, the interchange of the order of integration and Holder’s inequality. Note that there is nothing to check in the case of double sums. Thus if \(a_{ij} \geq 0\), it is always the case that

\[
\left( \sum_j \left( \sum_i a_{ij} \right)^p \right)^{1/p} \leq \sum_i \left( \sum_j a_{ij}^p \right)^{1/p}
\]

because the integrals in this case are just sums and \((i,j) \to a_{ij}\) is measurable.

The \(L^p\) spaces have many important properties.

### 22.2 Density Considerations

**Theorem 22.2.1** Let \(p \geq 1\) and let \((\Omega,\mathcal{S},\mu)\) be a measure space. Then the simple functions are dense in \(L^p(\Omega)\).

**Proof:** Recall that a function, \(f\), having values in \(\mathbb{R}\) can be written in the form \(f = f^+ - f^-\) where \(f^+ = \max(0, f)\), \(f^- = \max(0, -f)\).

Therefore, an arbitrary complex valued function, \(f\) is of the form

\[
f = \text{Re } f^+ - \text{Re } f^- + i (\text{Im } f^+ - \text{Im } f^-).
\]

If each of these nonnegative functions is approximated by a simple function, it follows \(f\) is also approximated by a simple function. Therefore, there is no loss of generality in assuming at the outset that \(f \geq 0\).

Since \(f\) is measurable, Theorem \(\mathbb{M}\) implies there is an increasing sequence of simple functions, \(\{s_n\}\), converging pointwise to \(f(x)\). Now

\[
|f(x) - s_n(x)| \leq |f(x)|.
\]

By the Dominated Convergence theorem,

\[
0 = \lim_{n \to \infty} \int |f(x) - s_n(x)|^p \, d\mu.
\]

Thus simple functions are dense in \(L^p\). □

Recall that for \(\Omega\) a topological space, \(C_c(\Omega)\) is the space of continuous functions with compact support in \(\Omega\). Also recall the following definition.
**Definition 22.2.2** Let $(\Omega, \mathcal{S}, \mu)$ be a measure space and suppose $(\Omega, \tau)$ is also a topological space. Then $(\Omega, \mathcal{S}, \mu)$ is called a regular measure space if the $\sigma$ algebra of Borel sets is contained in $\mathcal{S}$ and for all $E \in \mathcal{S}$,

$$\mu(E) = \inf \{ \mu(V) : V \supseteq E \text{ and } V \text{ open} \}$$

and if $\mu(E) < \infty$,

$$\mu(E) = \sup \{ \mu(K) : K \subseteq E \text{ and } K \text{ is compact} \}$$

and $\mu(K) < \infty$ for any compact set, $K$.

For example Lebesgue measure is an example of such a measure. More generally these measures are often referred to as Radon measures. The following useful lemma is stated here for convenience. It is Theorem 24.1.3 on Page 505.

**Lemma 22.2.3** Let $\Omega$ be a metric space in which the closed balls are compact and let $K$ be a compact subset of $\Omega$, an open set. Then there exists a continuous function $f : \Omega \to [0, 1]$ such that $f(x) = 1$ for all $x \in K$ and $\text{sp}(f)$ is a compact subset of $\text{V}$. That is, $K < f < \text{V}$.

It is not necessary to be in a metric space to do this. You can accomplish the same thing using Urysohn’s lemma.

**Theorem 22.2.4** Let $(\Omega, \mathcal{S}, \mu)$ be a regular measure space as in Definition 22.2.2 where the conclusion of Lemma 22.2.3 holds. Then $C_c(\Omega)$ is dense in $L^p(\Omega)$.

**Proof:** First consider a measurable set, $E$ where $\mu(E) < \infty$. Let $K \subseteq E \subseteq V$ where $\mu(V \setminus K) < \varepsilon$. Now let $K < h < V$. Then

$$\int |h - 1_E|^p \, d\mu \leq \int |1_V \setminus K|^p \, d\mu = \mu(V \setminus K) < \varepsilon.$$ 

It follows that for each $s$ a simple function in $L^p(\Omega)$, there exists $h \in C_c(\Omega)$ such that $\|s - h\|_p < \varepsilon$. This is because if

$$s(x) = \sum_{i=1}^m c_i \chi_{E_i}(x)$$

is a simple function in $L^p$ where the $c_i$ are the distinct nonzero values of $s$ each $\mu(E_i) < \infty$ since otherwise $s \notin L^p$ due to the inequality

$$\int |s|^p \, d\mu \geq |c_i|^p \mu(E_i).$$

By Theorem 24.2.1 simple functions are dense in $L^p(\Omega) \blacksquare$

### 22.3 Separability

**Theorem 22.3.1** For $p \geq 1$ and $\mu$ a Radon measure, $L^p(\mathbb{R}^n, \mu)$ is separable. Recall this means there exists a countable set, $\mathcal{D}$, such that if $f \in L^p(\mathbb{R}^n, \mu)$ and $\varepsilon > 0$, there exists $g \in \mathcal{D}$ such that $\|f - g\|_p < \varepsilon$.

**Proof:** Let $Q$ be all functions of the form $c\chi_{[a, b]}$ where

$$[a, b] = [a_1, b_1) \times [a_2, b_2) \times \cdots \times [a_n, b_n),$$

and both $a_i$, $b_i$ are rational, while $c$ has rational real and imaginary parts. Let $\mathcal{D}$ be the set of all finite sums of functions in $Q$. Thus, $\mathcal{D}$ is countable. In fact $\mathcal{D}$ is dense in $L^p(\mathbb{R}^n, \mu)$, $\|f - g\|_p$ is arbitrary, Theorem 24.2.1 implies
there exists \( g \in C_c(\mathbb{R}^n) \) such that \( ||f - g||_p \leq \frac{\varepsilon}{2} \) and then there would exist \( h \in C_c(\mathbb{R}^n) \) such that \( ||h - g||_p \leq \frac{\varepsilon}{2} \). By the triangle inequality,

\[
||f - h||_p \leq ||h - g||_p + ||g - f||_p < \varepsilon.
\]

Therefore, assume at the outset that \( f \in C_c(\mathbb{R}^n) \).

Let \( \mathcal{P}_m \) consist of all sets of the form \( [a, b] = \prod_{i=1}^n [a_i, b_i] \) where \( a_i = j2^{-m} \) and \( b_i = (j + 1)2^{-m} \) for \( j \) an integer. Thus \( \mathcal{P}_m \) consists of a tiling of \( \mathbb{R}^n \) into half open rectangles having diameters \( 2^{-m}n^\frac{1}{2} \). There are countably many of these rectangles; so, let \( \mathcal{P}_m = \{[a_i, b_i]\}_{i=1}^\infty \) and \( \mathbb{R}^n = \bigcup_{i=1}^\infty [a_i, b_i] \). Let \( \mathcal{C}_i^m \) be complex numbers with rational real and imaginary parts satisfying

\[
|f(a_i) - c_i^m| < 2^{-m},
\]

\[
|c_i^m| \leq |f(a_i)|. \tag{22.10}
\]

Let

\[
s_m(x) = \sum_{i=1}^\infty c_i^m \mathcal{X}_{[a_i, b_i]}(x).
\]

Since \( f(a_i) = 0 \) except for finitely many values of \( i \), the above is a finite sum. Then \( s_m \in D \). If \( s_m \) converges uniformly to \( f \) then it follows \( ||s_m - f||_p \to 0 \) because \( |s_m| \leq |f| \) and so

\[
||s_m - f||_p = \left( \int |s_m - f|^p \, d\mu \right)^{1/p} \leq \left( \int_{\text{spt}(f)} |s_m - f|^p \, d\mu \right)^{1/p} \leq [\varepsilon m_n(\text{spt}(f))]^{1/p}
\]

whenever \( m \) is large enough.

Since \( f \in C_c(\mathbb{R}^n) \) it follows that \( f \) is uniformly continuous and so given \( \varepsilon > 0 \) there exists \( \delta > 0 \) such that \( |x - y| < \delta, |f(x) - f(y)| < \varepsilon/2 \). Now let \( m \) be large enough that every box in \( \mathcal{P}_m \) has diameter less than \( \delta \) and also that \( 2^{-m} < \varepsilon/2 \). Then if \( [a_i, b_i] \) is one of these boxes of \( \mathcal{P}_m \), and \( x \in [a_i, b_i] \),

\[
|f(x) - f(a_i)| < \varepsilon/2
\]

and

\[
|f(a_i) - c_i^m| < 2^{-m} < \varepsilon/2.
\]

Therefore, using the triangle inequality, it follows that

\[
|f(x) - c_i^m| = |s_m(x) - f(x)| < \varepsilon
\]

and since \( x \) is arbitrary, this establishes uniform convergence. \( \blacksquare \)

**Corollary 22.3.2** Let \( \Omega \) be any \( \mu \) measurable subset of \( \mathbb{R}^n \) and let \( \mu \) be a Radon measure. Then \( L^p(\Omega, \mu) \) is separable. Here the \( \sigma \) algebra of measurable sets will consist of all intersections of measurable sets with \( \Omega \) and the measure will be \( \mu \) restricted to these sets.

**Proof:** Let \( \overline{D} \) be the restrictions of \( D \) to \( \Omega \). If \( f \in L^p(\Omega) \), let \( F \) be the zero extension of \( f \) to all of \( \mathbb{R}^n \). Let \( \varepsilon > 0 \) be given. By Theorem 22.3.1 there exists \( s \in D \) such that \( ||f - s||_p < \varepsilon \). Thus

\[
||s - f||_{L^p(\Omega, \mu)} \leq ||s - F||_{L^p(\mathbb{R}^n, \mu)} < \varepsilon
\]

and so the countable set \( \overline{D} \) is dense in \( L^p(\Omega) \).
22.4 Continuity Of Translation

**Definition 22.4.1** Let \( f \) be a function defined on \( U \subseteq \mathbb{R}^n \) and let \( w \in \mathbb{R}^n \). Then \( f_w \) will be the function defined on \( w + U \) by
\[
f_w(x) = f(x - w).
\]

**Theorem 22.4.2** (Continuity of translation in \( L^p \)) Let \( f \in L^p(\mathbb{R}^n) \) with the measure being Lebesgue measure. Then
\[
\lim_{||w|| \to 0} ||f_w - f||_p = 0.
\]

**Proof:** Let \( \varepsilon > 0 \) be given and let \( g \in C_c(\mathbb{R}^n) \) with \( ||g - f||_p < \frac{\varepsilon}{3} \). Since Lebesgue measure is translation invariant \( (m_n(w + E) = m_n(E)) \),
\[
||g_w - f_w||_p = ||g - f||_p < \frac{\varepsilon}{3}.
\]

You can see this from looking at simple functions and passing to the limit or you could use the change of variables formula to verify it.

Therefore
\[
||f - f_w||_p \leq ||f - g||_p + ||g - g_w||_p + ||g_w - f_w|| < \frac{2\varepsilon}{3} + ||g - g_w||_p. \tag{22.11}
\]

But \( \lim_{|w| \to 0} g_w(x) = g(x) \) uniformly in \( x \) because \( g \) is uniformly continuous. Now let \( B \) be a large ball containing \( \text{spt}(g) \) and let \( \delta_1 \) be small enough that \( B(x, \delta) \subseteq B \) whenever \( x \in \text{spt}(g) \). If \( \varepsilon > 0 \) is given there exists \( \delta < \delta_1 \) such that if \( |w| < \delta \), it follows that \( |g(x - w) - g(x)| < \varepsilon/3 \left(1 + m_n(B)^{1/p}\right) \). Therefore,
\[
||g - g_w||_p = \left( \int_B |g(x) - g(x - w)|^p \, dm_n \right)^{1/p} \leq \frac{\varepsilon m_n(B)^{1/p}}{3 \left(1 + m_n(B)^{1/p}\right)} < \frac{\varepsilon}{3}.
\]

Therefore, whenever \( |w| < \delta \), it follows \( ||g - g_w||_p < \frac{\varepsilon}{3} \) and so from 22.11 \( ||f - f_w||_p < \varepsilon \). This proves the theorem.

22.5 Mollifiers And Density Of Smooth Functions

**Definition 22.5.1** Let \( U \) be an open subset of \( \mathbb{R}^n \). \( C_c^\infty(U) \) is the vector space of all infinitely differentiable functions which equal zero for all \( x \) outside of some compact set contained in \( U \). Similarly, \( C_c^m(U) \) is the vector space of all functions which are \( m \) times continuously differentiable and whose support is a compact subset of \( U \).

**Example 22.5.2** Let \( U = B(z, 2r) \)
\[
\psi(x) = \begin{cases} 
\exp \left[ \left( |x - z|^2 - r^2 \right)^{-1} \right] & \text{if } |x - z| < r, \\
0 & \text{if } |x - z| \geq r.
\end{cases}
\]

Then a little work shows \( \psi \in C_c^\infty(U) \). The following also is easily obtained.

**Lemma 22.5.3** Let \( U \) be any open set. Then \( C_c^\infty(U) \neq \emptyset \).
Theorem 22.5.7 Let $f \in L^1_{\text{loc}}(\mathbb{R}^n, \mu)$ and $g \in C^\infty_c(\mathbb{R}^n)$. Then $f \ast g$ is an infinitely differentiable function. Here $\mu$ is a Radon measure on $\mathbb{R}^n$.

Proof: Consider the difference quotient for calculating a partial derivative of $f \ast g$.

$$
\frac{f \ast g(x + te_j) - f \ast g(x)}{t} = \int f(y) g(x + te_j - y) - g(x - y) \, d\mu(y).
$$

Using the fact that $g \in C^\infty_c(\mathbb{R}^n)$, the quotient,

$$
g(x + te_j - y) - g(x - y),
$$

is uniformly bounded. To see this easily, use Theorem 22.4.2 on Page 498 to get the existence of a constant, $M$ depending on

$$
\max \{ ||Dg(x)|| : x \in \mathbb{R}^n \}
$$

such that

$$
|g(x + te_j - y) - g(x - y)| \leq M |t|
$$

for any choice of $x$ and $y$. Therefore, there exists a dominating function for the integrand of the above integral which is of the form $C |f(y)| \mathcal{X}_K$ where $K$ is a compact set depending on the support of $g$. It follows the limit of the difference quotient above passes inside the integral as $t \to 0$ and

$$
\frac{\partial}{\partial x_j} (f \ast g)(x) = \int f(y) \frac{\partial}{\partial x_j} g(x - y) \, d\mu(y).
$$

Now letting $\frac{\partial}{\partial x_j} g$ play the role of $g$ in the above argument, partial derivatives of all orders exist. A similar use of the dominated convergence theorem shows all these partial derivatives are also continuous. This proves the lemma.

Recall also the following partition of unity result. It was proved earlier. See Corollary 22.4.3 on Page 498.
Theorem 22.5.8 Let $K$ be a compact set in $\mathbb{R}^n$ and let $\{U_i\}_{i=1}^{\infty}$ be an open cover of $K$. Then there exist functions, $\psi_k \in C_c^\infty(U_i)$ such that $\psi_i < U_i$ and for all $x \in K$,
\[ \sum_{i=1}^{\infty} \psi_i(x) = 1. \]

If $K_1$ is a compact subset of $U_1$ there exist such functions such that also $\psi_1(x) = 1$ for all $x \in K_1$.

Note that in the last conclusion of above corollary, the set $U_1$ could be replaced with $U_i$ for any fixed $i$ by simply renumbering.

Theorem 22.5.9 For each $p \geq 1$, $C_c^\infty(\mathbb{R}^n)$ is dense in $L^p(\mathbb{R}^n)$. Here the measure is Lebesgue measure.

Proof: Let $f \in L^p(\mathbb{R}^n)$ and let $\varepsilon > 0$ be given. Choose $g \in C_c(\mathbb{R}^n)$ such that $\|f - g\|_p < \frac{\varepsilon}{2}$. This can be done by using Theorem 22.7.3. Now let
\[ g_m(x) = g \ast \psi_m(x) \equiv \int g(x - y) \psi_m(y) \, dm_n(y) = \int g(y) \psi_m(x - y) \, dm_n(y) \]
where $\{\psi_m\}$ is a mollifier. It follows from Lemma 22.5.7 $g_m \in C_c^\infty(\mathbb{R}^n)$. It vanishes if $x \notin \text{spt}(g) + B(0, \frac{1}{m})$.

\[ \|g - g_m\|_p = \left( \int |g(x) - \int g(x - y) \psi_m(y) \, dm_n(y)|^p \, dm_n(x) \right)^{\frac{1}{p}} \]
\[ \leq \left( \int \left( \int |g(x) - g(x - y)| \psi_m(y) \, dm_n(y) \right|^p \, dm_n(x) \right)^{\frac{1}{p}} \]
\[ \leq \int \left( \int |g(x) - g(x - y)|^p \, dm_n(x) \right)^{\frac{1}{p}} \psi_m(y) \, dm_n(y) \]
\[ = \int_{B(0, \frac{1}{m})} \|g - g\|_p \psi_m(y) \, dm_n(y) < \frac{\varepsilon}{2} \]
whenever $m$ is large enough thanks to the uniform continuity of $g$. Theorem 22.7.12 was used to obtain the third inequality. There is no measurability problem because the function
\[ (x, y) \rightarrow |g(x) - g(x - y)| \psi_m(y) \]
is continuous. Thus when $m$ is large enough,
\[ \|f - g_m\|_p \leq \|f - g\|_p + \|g - g_m\|_p < \frac{\varepsilon}{2} + \varepsilon = \varepsilon. \]

This proves the theorem.

This is a very remarkable result. Functions in $L^p(\mathbb{R}^n)$ don’t need to be continuous anywhere and yet every such function is very close in the $L^p$ norm to one which is infinitely differentiable having compact support. The same result holds for $L^p(U)$ for $U$ an open set. This is the next corollary.

Corollary 22.5.10 Let $U$ be an open set. For each $p \geq 1$, $C_c^\infty(U)$ is dense in $L^p(U)$. Here the measure is Lebesgue measure.

Proof: Let $f \in L^p(U)$ and let $\varepsilon > 0$ be given. Choose $g \in C_c(U)$ such that $\|f - g\|_p < \frac{\varepsilon}{2}$. This is possible because Lebesgue measure restricted to the open set, $U$ is regular. Thus the existence of such a $g$ follows from Theorem 22.7.3. Now let
\[ g_m(x) = g \ast \psi_m(x) \equiv \int g(x - y) \psi_m(y) \, dm_n(y) = \int g(y) \psi_m(x - y) \, dm_n(y) \]
where \( \{ \psi_m \} \) is a mollifier. It follows from Lemma 22.5.7 that \( g_m \in C_c^\infty(U) \) for all \( m \) sufficiently large. It vanishes if \( x \notin \text{spt}(g) + B(0, \frac{1}{m}) \). Then

\[
\|g - g_m\|_p \leq \left( \int (|g(x) - g(x - y)| \psi_m(y)dm_n(y))^p dm_n(x) \right)^{\frac{1}{p}} 
\leq \left( \int (|g(x) - g(x - y)| \psi_m(y)dm_n(y))^p dm_n(x) \right)^{\frac{1}{p}} 
\leq \int (|g(x) - g(x - y)|^p dm_n(x))^{\frac{1}{p}} \psi_m(y)dm_n(y) 
= \int_{B(0, \frac{1}{m})} \|g - g_y\|_p \psi_m(y)dm_n(y) < \frac{\varepsilon}{2}
\]

whenever \( m \) is large enough thanks to uniform continuity of \( g \). Theorem 22.1.12 was used to obtain the third inequality. There is no measurability problem because the function \( (x, y) \mapsto |g(x) - g(x - y)| \psi_m(y) \) is continuous. Thus when \( m \) is large enough,

\[
\|f - g_m\|_p \leq ||f - g||_p + ||g - g_m||_p < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\]

This proves the corollary.

Another thing should probably be mentioned. If you have had a course in complex analysis, you may be wondering whether these infinitely differentiable functions having compact support have anything to do with analytic functions which also have infinitely many derivatives. The answer is no! Recall that if an analytic function has a limit point in the set of zeros then it is identically equal to zero. Thus these functions in \( C_c^\infty(\mathbb{R}^n) \) are not analytic. This is a strictly real analysis phenomenon and has absolutely nothing to do with the theory of functions of a complex variable.

### 22.6 Fundamental Theorem Of Calculus For Radon Measures

In this section the Besicovitch covering theorem will be used to give a differentiation theorem for general Radon measures. In what follows, \( \mu \) will be a Radon measure,

\[
Z \equiv \{ x \in \mathbb{R}^n : \mu(B(x, r)) = 0 \text{ for some } r > 0 \},
\]

**Lemma 22.6.1** \( Z \) is measurable and \( \mu(Z) = 0 \).

**Proof:** For each \( x \in Z \), there exists a ball \( B(x, r) \) with \( \mu(B(x, r)) = 0 \). Let \( \mathcal{C} \) be the collection of these balls. Since \( \mathbb{R}^n \) has a countable basis, a countable subset, \( \tilde{\mathcal{C}} \), of \( \mathcal{C} \) also covers \( Z \). Let

\[
\tilde{\mathcal{C}} = \{ B_i \}_{i=1}^\infty.
\]

Then letting \( \overline{\mu} \) denote the outer measure determined by \( \mu \),

\[
\overline{\mu}(Z) \leq \sum_{i=1}^\infty \overline{\mu}(B_i) = \sum_{i=1}^\infty \mu(B_i) = 0
\]

Therefore, \( Z \) is measurable and has measure zero as claimed. \( \blacksquare \)

Let \( Mf : \mathbb{R}^n \rightarrow [0, \infty] \) by

\[
Mf(x) = \begin{cases} 
\sup_{r \leq 1} \frac{1}{\mu(B(x, r))} \int_{B(x, r)} |f|d\mu & \text{if } x \notin Z \\
0 & \text{if } x \in Z
\end{cases}
\]
Theorem 22.6.2 Let $\mu$ be a Radon measure and let $f \in L^1(\mathbb{R}^n, \mu)$. Then for a.e. $x$,

$$\lim_{r \to 0} \frac{1}{\mu(B(x, r))} \int_{B(x, r)} |f(y) - f(x)| \, d\mu(y) = 0$$

**Proof:** First consider the following claim which is a weak type estimate of the same sort used when differentiating with respect to Lebesgue measure.

**Claim 1:** The following inequality holds for $N_n$ the constant of the Besicovitch covering theorem.

$$\overline{\mu}([Mf > \varepsilon]) \leq N_n \varepsilon^{-1} ||f||_1$$

**Proof:** First note $[Mf > \varepsilon] \cap Z = \emptyset$ and without loss of generality, you can assume $\overline{\mu}([Mf > \varepsilon]) > 0$. Next, for each $x \in [Mf > \varepsilon]$ there exists a ball $B_x = B(x, r_x)$ with $r_x \leq 1$ and

$$\mu(B_x)^{-1} \int_{B(x, r_x)} |f| \, d\mu > \varepsilon.$$

Let $F$ be this collection of balls so that $[Mf > \varepsilon]$ is the set of centers of balls of $F$. By the Besicovitch covering theorem,

$$[Mf > \varepsilon] \subseteq \bigcup_{i=1}^{N_n} \{B : B \in G_i\}$$

where $G_i$ is a collection of disjoint balls of $F$. Now for some $i$, $\overline{\mu}([Mf > \varepsilon]) / N_n \leq \mu(\cup \{B : B \in G_i\})$ because if this is not so, then

$$\overline{\mu}([Mf > \varepsilon]) \leq \sum_{i=1}^{N_n} \mu(\cup \{B : B \in G_i\})$$

$$< \sum_{i=1}^{N_n} \overline{\mu}([Mf > \varepsilon]) = \overline{\mu}([Mf > \varepsilon]),$$

a contradiction. Therefore for this $i$,

$$\frac{\overline{\mu}([Mf > \varepsilon])}{N_n} \leq \mu(\cup \{B : B \in G_i\}) = \sum_{B \in G_i} \mu(B) \leq \varepsilon^{-1} \int_B |f| \, d\mu \leq \varepsilon^{-1} \int_{\mathbb{R}^n} |f| \, d\mu = \varepsilon^{-1} ||f||_1.$$ 

This shows Claim 1.

**Claim 2:** If $g$ is any continuous function defined on $\mathbb{R}^n$, then for $x \notin Z$,

$$\lim_{r \to 0} \frac{1}{\mu(B(x, r))} \int_{B(x, r)} |g(y) - g(x)| \, d\mu(y) = 0$$

and

$$\lim_{r \to 0} \frac{1}{\mu(B(x, r))} \int_{B(x, r)} g(y) \, d\mu(y) = g(x). \quad (22.12)$$

**Proof:** Since $g$ is continuous at $x$, whenever $r$ is small enough,

$$\frac{1}{\mu(B(x, r))} \int_{B(x, r)} |g(y) - g(x)| \, d\mu(y) \leq \frac{1}{\mu(B(x, r))} \int_{B(x, r)} \varepsilon \, d\mu(y) = \varepsilon.$$

(22.12) follows from the above and the triangle inequality. This proves the claim.

Now let $g \in C_c(\mathbb{R}^n)$ and $x \notin Z$. Then from the above observations about continuous functions,

$$\overline{\mu} \left( \left\{ x \notin Z : \limsup_{r \to 0} \frac{1}{\mu(B(x, r))} \int_{B(x, r)} |f(y) - f(x)| \, d\mu(y) > \varepsilon \right\} \right) \quad (22.13)$$
22.6. FUNDAMENTAL THEOREM OF CALCULUS FOR RADON MEASURES

\[ \begin{align*}
\leq \ & \mathbb{P}\left( \left\{ x \notin Z : \limsup_{r \to 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r)} |f(y) - g(y)| \, d\mu(y) > \frac{\varepsilon}{2} \right\} \right) \\
& + \mathbb{P}\left( \left\{ x \notin Z : |g(x) - f(x)| > \frac{\varepsilon}{2} \right\} \right).
\end{align*} \]

(22.14)

Now
\[ \int_{\{ |f-g| > \frac{\varepsilon}{2} \}} |f-g| \, d\mu \geq \frac{\varepsilon}{2} \mathbb{P}\left( \left\{ |f-g| > \frac{\varepsilon}{2} \right\} \right), \]

and so from Claim 1 and hence 22.13 is dominated by
\[ \left( \frac{2}{\varepsilon} + \frac{N_n}{\varepsilon} \right) ||f-g||_{L^1(\mathbb{R}^n, \mu)}. \]

But by regularity of Radon measures, \( C_c(\mathbb{R}^n) \) is dense in \( L^1(\mathbb{R}^n, \mu) \) (See Theorem 22.4.11 for somewhat more than is needed.), and so since \( g \) in the above is arbitrary, this shows 22.13 equals 0. Now
\[ \begin{align*}
& \mathbb{P}\left( \left\{ x \notin Z : \limsup_{r \to 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r)} |f(y) - f(x)| \, d\mu(y) > 0 \right\} \right) \\
\leq \ & \sum_{k=1}^{\infty} \mathbb{P}\left( \left\{ x \notin Z : \limsup_{r \to 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r)} |f(y) - f(x)| \, d\mu(y) > \frac{1}{k} \right\} \right) = 0
\end{align*} \]

By completeness of \( \mu \) this implies
\[ \left\{ x \notin Z : \limsup_{r \to 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r)} |f(y) - f(x)| \, d\mu(y) > 0 \right\} \]
is a set of \( \mu \) measure zero. ■

The following corollary is the main result referred to as the Lebesgue Besicovitch Differentiation theorem. Recall \( L^1_{loc}(\mathbb{R}^n, \mu) \) refers to functions \( f \) which are measurable and also \( f\mathcal{X}_K \in L^1(\mathbb{R}^n, \mu) \) for any \( K \) compact.

**Corollary 22.6.3** If \( f \in L^1_{loc}(\mathbb{R}^n, \mu) \), then for a.e. \( x \notin Z \),
\[ \lim_{r \to 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r)} |f(y) - f(x)| \, d\mu(y) = 0. \] (22.15)

**Proof:** If \( f \) is replaced by \( f\mathcal{X}_{B(0,k)} \) then the conclusion 22.13 holds for all \( x \notin F_k \) where \( F_k \) is a set of \( \mu \) measure 0. Letting \( k = 1, 2, \ldots, \) and \( F \equiv \bigcup_{k=1}^{\infty} F_k \), it follows that \( F \) is a set of measure zero and for any \( x \notin F \), and \( k \in \{1, 2, \ldots, \} \), 22.13 holds if \( f \) is replaced by \( f\mathcal{X}_{B(0,k)} \). Picking any such \( x \), and letting \( k > |x| + 1 \), this shows
\[ \lim_{r \to 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r)} |f(y) - f(x)| \, d\mu(y) = \lim_{r \to 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r)} |f\mathcal{X}_{B(0,k)}(y) - f\mathcal{X}_{B(0,k)}(x)| \, d\mu(y) = 0. \] ■

In case \( \mu \) is ordinary Lebesgue measure on \( \mathbb{R}^n \), the set \( Z = \emptyset \) and we obtain the usual Lebesgue fundamental theorem of calculus.

**Corollary 22.6.4** If \( f \in L^1_{loc}(\mathbb{R}^n, m_n) \), then for a.e. \( x \),
\[ \lim_{r \to 0} \frac{1}{m_n(B(x,r))} \int_{B(x,r)} |f(y) - f(x)| \, dm_n(y) = 0. \] (22.16)

In particular, for a.e. \( x \),
\[ \lim_{r \to 0} \frac{1}{m_n(B(x,r))} \int_{B(x,r)} f(y) \, dm_n(y) = f(x). \]
22.7 Exercises

1. Let $E$ be a Lebesgue measurable set in $\mathbb{R}$. Suppose $m(E) > 0$. Consider the set

$$E - E = \{x - y : x \in E, y \in E\}.$$ 

Show that $E - E$ contains an interval. Hint: Let

$$f(x) = \int \chi_E (t) \chi_E (x + t) dt.$$ 

Note $f$ is continuous at 0 and $f(0) > 0$ and use continuity of translation in $L^p$.

2. Establish the inequality $||fg||_r \leq ||f||_p ||g||_q$ whenever $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$.

3. Let $(\Omega, S, \mu)$ be counting measure on $\mathbb{N}$. Thus $\Omega = \mathbb{N}$ and $S = \mathcal{P} (\mathbb{N})$ with $\mu(S) =$ number of things in $S$. Let $1 \leq p \leq q$. Show that in this case,

$$L^1 (\mathbb{N}) \subseteq L^p (\mathbb{N}) \subseteq L^q (\mathbb{N}).$$

Hint: This is real easy if you consider what $\int_{\Omega} f d\mu$ equals. How are the norms related?

4. Consider the function, $f(x, y) = \frac{x^{p-1}}{py} + \frac{y^{q-1}}{qx}$ for $x, y > 0$ and $\frac{1}{p} + \frac{1}{q} = 1$. Show directly that $f(x, y) \geq 1$ for all such $x, y$ and show this implies $xy \leq \frac{x^p}{p} + \frac{y^q}{q}$.

5. Give an example of a sequence of functions in $L^p (\mathbb{R})$ which converges to zero in $L^p$ but does not converge pointwise to 0. Does this contradict the proof of the theorem that $L^p$ is complete?

6. Let $K$ be a bounded subset of $L^p (\mathbb{R}^n)$ and suppose that there exists $G$ such that $\overline{G}$ is compact with

$$\int_{\mathbb{R}^n \setminus \overline{G}} |u(x)|^p \, dx < \varepsilon^p$$

and for all $\varepsilon > 0$, there exist a $\delta > 0$ and such that if $|h| < \delta$, then

$$\int |u(x + h) - u(x)|^p \, dx < \varepsilon^p$$

for all $u \in K$. Show that $K$ is precompact in $L^p (\mathbb{R}^n)$. Hint: Let $\phi_k$ be a mollifier and consider

$$K_k \equiv \{u * \phi_k : u \in K\}.$$ 

Verify the conditions of the Ascoli Arzela theorem for these functions defined on $\overline{G}$ and show there is an $\varepsilon$ net for each $\varepsilon > 0$. Can you modify this to let an arbitrary open set take the place of $\mathbb{R}^n$?

7. Let $(\Omega, d)$ be a metric space and suppose also that $(\Omega, S, \mu)$ is a regular measure space such that $\mu(\Omega) < \infty$ and let $f \in L^1 (\Omega)$ where $f$ has complex values. Show that for every $\varepsilon > 0$, there exists an open set of measure less than $\varepsilon$, denoted here by $V$ and a continuous function, $g$ defined on $\Omega$ such that $f = g$ on $V^C$. Thus, aside from a set of small measure, $f$ is continuous. If $|f(\omega)| \leq M$, show that it can be assumed that $|g(\omega)| \leq M$. This is called Lusin’s theorem. Hint: Use Theorems 22.3.1 and 22.4.1 to obtain a sequence of functions in $C_c (\Omega), \{g_n\}$ which converges pointwise a.e. to $f$ and then use Egoroff’s theorem to obtain a small set, $W$ of measure less than $\varepsilon/2$ such that convergence is uniform on $W^C$. Now let $F$ be a closed subset of $W^C$ such that $\mu(W^C \setminus F) < \varepsilon/2$. Let $V = F^C$. Thus $\mu(V) < \varepsilon$ and on $F = V^C$, the convergence of $\{g_n\}$ is uniform showing that the restriction of $f$ to $V^C$ is continuous. Now use the Tietze extension theorem.
8. Let $\phi_m \in C_c^\infty(\mathbb{R}^n)$, $\phi_m(x) \geq 0$, and $\int_{\mathbb{R}^n} \phi_m(y)dy = 1$ with

$$\lim_{m \to \infty} \sup \{ |x| : x \in \text{spt}(\phi_m) \} = 0.$$ 

Show if $f \in L^p(\mathbb{R}^n)$, $\lim_{m \to \infty} f * \phi_m = f$ in $L^p(\mathbb{R}^n)$.

9. Let $\phi : \mathbb{R} \to \mathbb{R}$ be convex. This means

$$\phi(\lambda x + (1 - \lambda)y) \leq \lambda \phi(x) + (1 - \lambda)\phi(y)$$

whenever $\lambda \in [0, 1]$. Verify that if $x < y < z$, then $\frac{\phi(y) - \phi(x)}{y - x} \leq \frac{\phi(z) - \phi(x)}{z - x}$ and that $\frac{\phi(z) - \phi(x)}{z - x} \leq \frac{\phi(z) - \phi(y)}{z - y}$. Show if $s \in \mathbb{R}$ there exists $\lambda$ such that $\phi(s) \leq \phi(t) + \lambda(s - t)$ for all $t$. Show that if $\phi$ is convex, then $\phi$ is continuous.

10. Prove Jensen’s inequality. If $\phi : \mathbb{R} \to \mathbb{R}$ is convex, $\mu(\Omega) = 1$, and $f : \Omega \to \mathbb{R}$ is in $L^1(\Omega)$, then $\phi(\int_{\Omega} f du) \leq \int_{\Omega} \phi(f)du$. Hint: Let $s = \int_{\Omega} f du$ and use Problem 6.

11. Let $\frac{1}{p} + \frac{1}{q} = 1$, $p > 1$, let $f \in L^p(\mathbb{R})$, $g \in L^q(\mathbb{R})$. Show $f * g$ is uniformly continuous on $\mathbb{R}$ and $\|(f * g)(x)\| \leq \|f\|_p \|g\|_q$. Hint: You need to consider why $f * g$ exists and then this follows from the definition of convolution and continuity of translation in $L^p$.

12. $B(p, q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx$, $\Gamma(p) = \int_0^\infty e^{-t}t^{p-1}dt$ for $p, q > 0$. The first of these is called the beta function, while the second is the gamma function. Show a.) $\Gamma(p+1) = p\Gamma(p)$; b.) $\Gamma(p)\Gamma(q) = B(p, q)\Gamma(p+q)$.

13. Let $f \in C_c((0, \infty))$ and define $F(x) = \frac{1}{2} \int_0^x f(t)dt$. Show

$$\|F\|_{L^p(0, \infty)} \leq \frac{p}{p-1}\|f\|_{L^p(0, \infty)}$$

whenever $p > 1$.

Hint: Argue there is no loss of generality in assuming $f \geq 0$ and then assume this is so. Integrate $\int_0^\infty |F(x)|^pdx$ by parts as follows:

$$\int_0^\infty F^pdx = \int_0^\infty xF^{p-1}F'dx.$$ 

Now show $xF' = f - F$ and use this in the last integral. Complete the argument by using Holder’s inequality and $p - 1 = p/q$.

14. Suppose $f \in L^p((0, \infty))$, $p > 1$, and $f$ not necessarily in $C_c((0, \infty))$. Show that $F(x) = \frac{1}{2} \int_0^x f(t)dt$ still makes sense for each $x > 0$. Show the inequality of Problem 10 is still valid. This inequality is called Hardy’s inequality. Hint: To show this, use the above inequality along with the density of $C_c(0, \infty)$ in $L^p(0, \infty)$.

15. Suppose $f, g \geq 0$. When does equality hold in Holder’s inequality?

16. Show the Vitali Convergence theorem implies the Dominated Convergence theorem for finite measure spaces but there exist examples where the Vitali convergence theorem works and the dominated convergence theorem does not.

17. Suppose $\mu(\Omega) < \infty$, $\{f_n\} \subseteq L^1(\Omega)$, and

$$\int_\Omega h(|f_n|)d\mu < C$$

for all $n$ where $h$ is a continuous, nonnegative function satisfying

$$\lim_{t \to \infty} \frac{h(t)}{t} = \infty.$$ 

Show $\{f_n\}$ is uniformly integrable. In applications, this often occurs in the form of a bound on $\|f_n\|_p$. 

18. ↑ Sometimes, especially in books on probability, a different definition of uniform integrability is used than that presented here. A set of functions, \( \mathcal{S} \), defined on a finite measure space, \((\Omega, \mathcal{S}, \mu)\) is said to be uniformly integrable if for all \( \varepsilon > 0 \) there exists \( \alpha > 0 \) such that for all \( f \in \mathcal{S} \),

\[
\int_{|f| \geq \alpha} |f| \, d\mu \leq \varepsilon.
\]

Show that this definition is equivalent to the definition of uniform integrability with the addition of the condition that there is a constant, \( C < \infty \) such that

\[
\int |f| \, d\mu \leq C
\]

for all \( f \in \mathcal{S} \).

19. \( f \in L^\infty(\Omega, \mu) \) if there exists a set of measure zero, \( E \), and a constant \( C < \infty \) such that \( |f(x)| \leq C \) for all \( x \notin E \).

\[
||f||_\infty \equiv \inf \{ C : |f(x)| \leq C \ a.e. \}.
\]

Show \( || \cdot ||_\infty \) is a norm on \( L^\infty(\Omega, \mu) \) provided \( f \) and \( g \) are identified if \( f(x) = g(x) \) a.e. Show \( L^\infty(\Omega, \mu) \) is complete. **Hint:** You might want to show that \( ||f|| > ||f||_\infty \) has measure zero so \( ||f||_\infty \) is the smallest number at least as large as \( |f(x)| \) for a.e. \( x \). Thus \( ||f||_\infty \) is one of the constants, \( C \) in the above.

20. Suppose \( f \in L^\infty \cap L^1 \). Show \( \lim_{p \to \infty} ||f||_{L^p} = ||f||_\infty \). **Hint:**

\[
(||f||_\infty - \varepsilon)^p \mu(||f| > ||f||_\infty - \varepsilon) \leq \int_{||f| > ||f||_\infty - \varepsilon} |f|^p \, d\mu \leq \int |f|^p \, d\mu
\]

Now raise both ends to the \( 1/p \) power and take lim inf and lim sup as \( p \to \infty \). You should get \( ||f||_\infty - \varepsilon \leq \lim \inf ||f||_p \leq \lim \sup ||f||_p \leq ||f||_\infty \)

21. Suppose \( \mu(\Omega) < \infty \). Show that if \( 1 \leq p < q \), then \( L^q(\Omega) \subseteq L^p(\Omega) \). **Hint** Use Holder’s inequality.

22. Show \( L^1(\mathbb{R}) \not\subseteq L^2(\mathbb{R}) \) and \( L^2(\mathbb{R}) \not\subseteq L^1(\mathbb{R}) \) if Lebesgue measure is used. **Hint:** Consider \( 1/\sqrt{x} \) and \( 1/x \).

23. Suppose that \( \theta \in [0,1] \) and \( r, s, q > 0 \) with

\[
\frac{1}{q} = \frac{\theta}{r} + \frac{1 - \theta}{s}.
\]

show that

\[
(\int |f|^q \, d\mu)^{1/q} \leq (\int |f|^r \, d\mu)^{1/r} (\int |f|^s \, d\mu)^{1/s})^{1 - \theta}.
\]

If \( q, r, s \geq 1 \) this says that

\[
||f||_q \leq ||f||_r^{\theta} ||f||_s^{1 - \theta}.
\]

Using this, show that

\[
\ln (||f||_q) \leq \theta \ln (||f||_r) + (1 - \theta) \ln (||f||_s).
\]

**Hint:**

\[
\int |f|^q \, d\mu = \int |f|^q \, f^{1/q-1} \, d\mu.
\]

Now note that \( 1 = \frac{\theta q}{r} + \frac{\theta(1-\theta)}{s} \) and use Holder’s inequality.
24. Suppose $f$ is a function in $L^1(\mathbb{R})$ and $f$ is infinitely differentiable. Is $f' \in L^1(\mathbb{R})$? **Hint:** What if $\phi \in C_c^{\infty}(0,1)$ and $f(x) = \phi(2^n (x - n))$ for $x \in (n, n + 1)$, $f(x) = 0$ if $x < 0$?

25. Let $(\mathbb{R}^n, \mathcal{F}, \mu)$ be a measure space with $\mu$ a Radon measure. That is, it is regular, $\mathcal{F}$ contains the Borel sets, $\mu$ is complete, and finite on compact sets. Let $A$ be a measurable set. Show that for a.e. $x \in A$,

$$1 = \lim_{r \to 0} \frac{\mu(A \cap B(x, r))}{\mu(B(x, r))}.$$ 

Such points are called “points of density”. **Hint:** The above quotient is nothing more than

$$\frac{1}{\mu(B(x, r))} \int_{B(x, r)} X_A(x) \, d\mu$$

Now consider Corollary 22.6.3.
Chapter 23

Representation Theorems

23.1 Basic Theory

As explained earlier, a normed linear space is a vector space $X$ with a norm. This is a map $\| \cdot \| : X \to [0, \infty)$ which satisfies the following axioms.

1. $\| x \| \geq 0$ and equals 0 if and only if $x = 0$

2. For $\alpha$ a scalar and $x$ a vector in $X$, $\| \alpha x \| = |\alpha| \| x \|$

3. $\| x + y \| \leq \| x \| + \| y \|$

Then, as discussed earlier, this is a metric space if $d(x, y) = \| x - y \|$. The field of scalars will be either $\mathbb{R}$ or $\mathbb{C}$, usually $\mathbb{C}$. Then as before, there is a definition of an inner product space. If $(X, \| \cdot \|)$ is complete, meaning that Cauchy sequences converge, then this is called a Banach space. A whole lot can be said about Banach spaces but not in this book. Here, a specialization will be considered which ties in well with the earlier material on inner product spaces.

Definition 23.1.1 Let $X$ be a vector space. An inner product is a mapping from $X \times X$ to $\mathbb{C}$ if $X$ is complex and from $X \times X$ to $\mathbb{R}$ if $X$ is real, denoted by $(x, y)$ which satisfies the following.

\begin{align*}
(x, x) &\geq 0, \quad (x, x) = 0 \text{ if and only if } x = 0, \\
(x, y) &= \overline{(y, x)}. \tag{23.1}
\end{align*}

For $a, b \in \mathbb{C}$ and $x, y, z \in X$,

\begin{align*}
(ax + by, z) &= a(x, z) + b(y, z). \tag{23.2}
\end{align*}

Note that (23.2) and (23.3) imply $(x, ay + bz) = \overline{a}(x, y) + \overline{b}(x, z)$. Such a vector space is called an inner product space.

The Cauchy Schwarz inequality is fundamental for the study of inner product spaces. The proof is identical to that given earlier in Theorem 11.4.4 on Page 253.

Theorem 23.1.2 (Cauchy Schwarz) In any inner product space

\begin{align*}
|\langle x, y \rangle| &\leq \| x \| \| y \|. \tag{23.3}
\end{align*}

Also, as earlier, the norm given by the inner product, really is a norm.

Proposition 23.1.3 For an inner product space, $\| x \| = (x, x)^{1/2}$ does specify a norm.

The following lemma is called the parallelogram identity. It was also discussed earlier.
Lemma 23.1.4 In an inner product space,

\[ \|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2. \]

The proof, a straightforward application of the inner product axioms, is left to the reader.

Lemma 23.1.5 For \( x \in H \), an inner product space,

\[ \|x\| = \sup_{\|y\| \leq 1} |(x, y)| \tag{23.4} \]

Proof: By the Cauchy Schwarz inequality, if \( x \neq 0 \),

\[ \|x\| \geq \sup_{\|y\| \leq 1} |(x, y)| \geq \left( x, \frac{x}{\|x\|} \right) = \|x\|. \]

It is obvious that (23.4) holds in the case that \( x = 0 \).

Definition 23.1.6 A Hilbert space is an inner product space which is complete. Thus a Hilbert space is a Banach space in which the norm comes from an inner product as described above. Often people use the symbol \( \| \cdot \| \) to denote the norm rather than \( || \cdot || \).

In Hilbert space, one can define a projection map onto closed convex nonempty sets.

Definition 23.1.7 A set, \( K \), is convex if whenever \( \lambda \in [0, 1] \) and \( x, y \in K \), \( \lambda x + (1 - \lambda)y \in K \).

This was done in the problems beginning with Problem 3 on Page 204.

Theorem 23.1.8 Let \( K \) be a closed convex nonempty subset of a Hilbert space, \( H \), and let \( x \in H \). Then there exists a unique point \( Px \in K \) such that \( \|Px - x\| \leq \|y - x\| \) for all \( y \in K \).

Corollary 23.1.9 Let \( K \) be a closed, convex, nonempty subset of a Hilbert space, \( H \), and let \( x \in H \). Then for \( z \in K \), \( z = Px \) if and only if

\[ \text{Re}(x - z, y - z) \leq 0 \tag{23.5} \]

for all \( y \in K \).

Here we present an easier version which will be sufficient for what is needed. First is a simple lemma which is interesting for its own sake.

Lemma 23.1.10 Suppose \( \text{Re}(w, y) = 0 \) for all \( y \in M \) a subspace of \( H \), an inner product space. Then \( (w, y) = 0 \).

Proof: Consider the following:

\[ (w, y) = \text{Re}(w, y) + i \text{Im}(w, y) \]

\[ (w, iy) = -i (w, y) = -i (\text{Re}(w, y) + i \text{Im}(w, y)) \]

and so \( \text{Im}(w, y) = \text{Re}(w, iy) = 0 \). Since \( \text{Im}(w, y) = 0 \) as well as \( \text{Re}(w, y) \), it follows that \( (w, y) = 0 \).

Theorem 23.1.11 Let \( H \) be a Hilbert space and let \( M \) be a closed subspace. Then if \( w \in H \), there exists a unique \( Pw \in H \) such that

\[ \|Pw - w\| \leq \|y - w\| \]

for all \( y \in M \). Thus \( Pw \) is the point of \( M \) closest to \( w \) and it is unique. Then \( z \in M \) equals \( Pw \) if and only if

\[ (w - z, y) = 0 \] for all \( y \in M \)

Also the map \( w \to Pw \) is a linear map which satisfies \( \|Pw\| \leq \|w\| \).
23.1. BASIC THEORY

**Proof:** Let \(\lambda \equiv \inf \{\|y-w\| : y \in M\}\). Let \(\{y_n\}\) be a minimizing sequence. Say \(\|y_n-w\| < \lambda + \frac{1}{n}\). Then, by the parallelogram identity,

\[
\frac{1}{2} \left(\left\|y_n-y_m\right\|^2 + \left\|y_n+y_m\right\|^2\right) \leq \frac{1}{4} \left(\left\|y_n+y_m\right\|^2 - \left\|y_n-y_m\right\|^2\right)
\]

and clearly the right side converges to 0 if \(m,n \to \infty\). Thus this is a Cauchy sequence and so it converges to some \(z \in M\) since \(M\) is closed. Then

\[
\lambda \leq \|z-w\| = \lim_{n \to \infty} \|y_n-w\| \leq \lambda
\]

and so \(Pw = z\) is a closest point. There can be only one closest point because if \(z_i\) works, then by the parallelogram identity again,

\[
\frac{1}{2} \left(\left\|z_1+z_2\right\|^2 + \left\|z_1-z_2\right\|^2\right) = \frac{1}{2} \left(\left\|z_1-w\right\|^2 + \left\|z_2-w\right\|^2\right) - \left\|z_1-z_2\right\|^2
\]

and so if these are different, then \(\frac{z_1+z_2}{2}\) is closer to \(w\) than \(\lambda\) contradicting the choice of \(z_i\).

Now for the characterization: For \(z \in M\), let \(y \in M\) consider \(z+t(y-z)\) for \(t \in \mathbb{R}\). Then

\[
\|w-(z+t(y-z))\|^2 = \|w-z\|^2 + t^2 \|y-z\|^2 - 2t \text{Re} \langle w-z, y-z \rangle
\]

(*)

Then for \(z\) to be the closest point to \(w\), one needs the above to be minimized when \(t = 0\). Taking a derivative, this requires that

\[
\text{Re} \langle w-z, y-z \rangle = 0
\]

for any \(y \in M\). But this is the same as saying that \(\text{Re} \langle w-z, y \rangle = 0\) for all \(y \in M\). By Lemma \(\mathbf{6.1}\), \(\langle w-z, y \rangle = 0\) for all \(y \in M\). Thus \(\langle w-z, y \rangle = 0\) if \(z = Pw\).

Conversely, if \(\langle w-z, y \rangle = 0\) for all \(y \in M\), then * shows that \(\|w-(z+t(y-z))\|^2\) achieves its minimum when \(t = 0\) for any \(y\). But a generic point of \(M\) is of the form \(z+t(y-z)\) and so \(z = Pw\).

As to \(P\) being linear, for \(y \in M\) arbitrary,

\[
0 = \langle \alpha w_1 + \beta w_2 - P(\alpha w_1 + \beta w_2), y \rangle
\]

Also,

\[
0 = \langle \alpha w_1 + \beta w_2 - (\alpha P(w_1) + \beta P(w_2)), y \rangle
\]

By uniqueness, \(P(\alpha w_1 + \beta w_2) = \alpha P(w_1) + \beta P(w_2)\).

Finally,

\[
\|w-Pw, Pw\| = 0, \quad \|Pw\|^2 = \langle w, Pw \rangle \leq \|w\| \|Pw\|
\]

which yields the desired estimate. \(\blacksquare\)

Note that the operator norm of \(P\) equals 1.

\[
\|P\| = \sup_{\|w\| \leq 1} \|Pw\| \leq \sup_{\|w\| \leq 1} \|w\| = 1
\]

Now pick \(w \in M\) with \(\|w\| = 1\). Then \(\|P\| \geq \|Pw\| = \|w\| = 1\).
Definition 23.1.12 If \( A : X \to Y \) is linear where \( X, Y \) are two normed linear spaces, then \( A \) is said to be in \( \mathcal{L}(X,Y) \) if and only if

\[
\|A\| = \sup_{\|x\| \leq 1} \|Ax\|_Y < \infty
\]

In case that \( Y = \mathbb{F} \) the field of scalars, equal to either \( \mathbb{R} \) or \( \mathbb{C} \), \( \mathcal{L}(X,\mathbb{F}) \) is known as the dual space, written here as \( X' \). Actually it is more often written as \( X^* \) but I prefer the former notation because the latter is sometimes used to denote a purely algebraic dual space, meaning only that its elements are linear maps into \( \mathbb{F} \) with no requirement of continuity. Doubtless there are drawbacks to my notation also.

Thus \( P \) the projection map in the above is in \( \mathcal{L}(H,H) \).

There is a general easy result about \( \mathcal{L}(X,Y) \) which follows. It says that these linear maps are continuous.

Theorem 23.1.13 Let \( X \) and \( Y \) be two normed linear spaces and let \( L : X \to Y \) be linear (\( L(ax + by) = aL(x) + bL(y) \) for \( a, b \) scalars and \( x, y \in X \)). The following are equivalent

\begin{enumerate} \item \( L \) is continuous at 0 \item \( L \) is continuous \item There exists \( K > 0 \) such that \( \|Lx\|_Y \leq K \|x\|_X \) for all \( x \in X \) (\( L \) is bounded). \end{enumerate}

Proof: \( \text{a.)} \Rightarrow \text{b.)} \) Let \( x_n \to x \). It is necessary to show that \( Lx_n \to Lx \). But \( (x_n - x) \to 0 \) and so from continuity at 0, it follows

\[ L(x_n - x) = Lx_n - Lx \to 0 \]

so \( Lx_n \to Lx \). This shows \( \text{a.)} \) implies \( \text{b.)} \).

\( \text{b.)} \Rightarrow \text{c.)} \) Since \( L \) is continuous, \( L \) is continuous at 0. Hence \( \|Lx\|_Y < 1 \) whenever \( \|x\|_X \leq \delta \) for some \( \delta \). Therefore, suppressing the subscript on the \( \| \| \),

\[ \|L\left(\frac{\delta x}{\|x\|}\right)\| \leq 1. \]

Hence

\[ \|Lx\| \leq \frac{1}{\delta} \|x\|. \]

\( \text{c.)} \Rightarrow \text{a.)} \) follows from the inequality given in \( \text{c.)} \).

The following theorem is called the Riesz representation theorem for the dual of a Hilbert space.

If \( z \in H \) then define an element \( f \in H' \) by the rule \( (x,z) \equiv f(x) \). It follows from the Cauchy Schwarz inequality and the properties of the inner product that \( f \in H' \). The Riesz representation theorem says that all elements of \( H' \) are of this form.

Theorem 23.1.14 Let \( H \) be a Hilbert space and let \( f \in H' \). Then there exists a unique \( z \in H \) such that

\[ f(x) = (x,z) \quad (23.8) \]

for all \( x \in H \).

Proof: Letting \( y, w \in H \) the assumption that \( f \) is linear implies

\[ f(yf(w) - f(y)w) = f(w)f(y) - f(y)f(w) = 0 \]

which shows that \( yf(w) - f(y)w \in f^{-1}(0) \), which is a closed subspace of \( H \) since \( f \) is continuous. If \( f^{-1}(0) = H \), then \( f \) is the zero map and \( z = 0 \) is the unique element of \( H \) which satisfies \( \text{23.8} \).

If \( f^{-1}(0) \neq H \), pick \( u \notin f^{-1}(0) \) and let \( w \equiv u - Pu \neq 0 \). Thus \( (y,w) = 0 \) for all \( y \in f^{-1}(0) \). In particular, let \( y = xf(w) - f(x)w \) where \( x \in H \) is arbitrary. Therefore,

\[ 0 = (f(w)x - f(x)w, w) = f(w)(x, w) - f(x)||w||^2. \]
Thus, solving for \(f(x)\) and using the properties of the inner product,

\[
f(x) = \langle x, \frac{\overline{f(w)}w}{\|w\|^2} \rangle
\]

Let \(z = \frac{\overline{f(w)}w}{\|w\|^2}\). This proves the existence of \(z\). If \(f(x) = (x, z_i)\) \(i = 1, 2\), for all \(x \in H\), then for all \(x \in H\), then \((x, z_1 - z_2) = 0\) which implies, upon taking \(x = z_1 - z_2\) that \(z_1 = z_2\). 

If \(R : H \to H^\prime\) is defined by \(Rx(y) \equiv (y, x)\), the Riesz representation theorem above states this map is onto. This map is called the Riesz map. It is routine to show \(R\) is conjugate linear and \(\|Rx\| = \|x\|\). In fact,

\[
R(\alpha x + \beta y)(u) = (\alpha u, \alpha x + \beta y) = \alpha (u, x) + \beta (u, y)
\]

\[
= \alpha Rx(u) + \beta Ry(u) = (\alpha Rx + \beta Ry)(u)
\]

so it is conjugate linear meaning it goes across plus signs and you factor out conjugates.

\[
\|Rx\| = \sup_{\|y\| \leq 1} |Rx(y)| = \sup_{\|y\| \leq 1} \|(y, x)\| = \|x\|
\]

### 23.2 Radon Nikodym Theorem

The Radon Nikodym theorem, is a representation theorem for one measure in terms of another. The approach given here is due to Von Neumann and depends on the Riesz representation theorem for Hilbert space, Theorem 23.1.1.

**Definition 23.2.1** Let \(\mu\) and \(\lambda\) be two measures defined on a \(\sigma\)-algebra \(S\), of subsets of a set, \(\Omega\). \(\lambda\) is absolutely continuous with respect to \(\mu\), written as \(\lambda \ll \mu\), if \(\lambda(E) = 0\) whenever \(\mu(E) = 0\).

It is not hard to think of examples which should be like this. For example, suppose one measure is volume and the other is mass. If the volume of something is zero, it is reasonable to expect the mass of it should also be equal to zero. In this case, there is a function called the density which is integrated over volume to obtain mass. The Radon Nikodym theorem is an abstract version of this notion. Essentially, it gives the existence of the density function.

**Theorem 23.2.2** (Radon Nikodym) Let \(\lambda\) and \(\mu\) be finite measures defined on a \(\sigma\)-algebra, \(S\), of subsets of \(\Omega\). Suppose \(\lambda \ll \mu\). Then there exists a unique \(f \in L^1(\Omega, \mu)\) such that \(f(x) \geq 0\) and

\[
\lambda(E) = \int_E f \, d\mu.
\]

If it is not necessarily the case that \(\lambda \ll \mu\), there are two measures, \(\lambda_\perp\) and \(\lambda_\|\) such that \(\lambda = \lambda_\perp + \lambda_\|\ll\mu\) and there exists a set of \(\mu\) measure zero, \(N\) such that for all \(E\) measureable, \(\lambda_\perp(E) = \lambda(E \cap N) = \lambda_\|E \cap N\). In this case the two measures, \(\lambda_\perp\) and \(\lambda_\|\) are unique and the representation of \(\lambda = \lambda_\perp + \lambda_\|\) is called the Lebesgue decomposition of \(\lambda\). The measure \(\lambda_\|\) is the absolutely continuous part of \(\lambda\) and \(\lambda_\perp\) is called the singular part of \(\lambda\). In words, \(\lambda_\perp\) is “supported” on a set of \(\mu\) measure zero while \(\lambda_\|\) is supported on the complement of this set.

**Proof:** Let \(\Lambda : L^2(\Omega, \mu + \lambda) \to \mathbb{C}\) be defined by

\[
\Lambda g = \int_\Omega g \, d\lambda.
\]

By Holder’s inequality,

\[
|\Lambda g| \leq \left(\int_\Omega 1^2 \, d\lambda \right)^{1/2} \left(\int_\Omega |g|^2 \, d(\lambda + \mu) \right)^{1/2} = \lambda(\Omega)^{1/2} \|g\|_2
\]
where \( \|g\|_2 \) is the \( L^2 \) norm of \( g \) taken with respect to \( \mu + \lambda \). Therefore, since \( \Lambda \) is bounded, it follows from Theorem 23.1.13 on Page 558 that \( \Lambda \in (L^2(\Omega, \mu + \lambda))^\prime \), the dual space \( L^2(\Omega, \mu + \lambda) \). By the Riesz representation theorem in Hilbert space, Theorem 23.1.14, there exists a unique \( h \in L^2(\Omega, \mu + \lambda) \) with

\[
\Lambda g = \int_\Omega g \, d\lambda = \int_\Omega hgd(\mu + \lambda).
\]  

(23.9)

The plan is to show \( h \) is real and nonnegative at least a.e. Therefore, consider the set where \( \text{Im} \, h \) is positive.

\[ E = \{ x \in \Omega : \text{Im} \, h(x) > 0 \}, \]

Now let \( g = \mathcal{X}_E \) and use 23.9 to get

\[ \lambda(E) = \int_E (\text{Re} \, h + i \, \text{Im} \, h)d(\mu + \lambda). \]

(23.10)

Since the left side of 23.10 is real, this shows

\[ 0 = \int_E (\text{Im} \, h) \, d(\mu + \lambda) \geq \int_{E_n} (\text{Im} \, h) \, d(\mu + \lambda) \geq \frac{1}{n} (\mu + \lambda) (E_n) \]

where

\[ E_n \equiv \left\{ x : \text{Im} \, h(x) \geq \frac{1}{n} \right\} \]

Thus \( (\mu + \lambda) (E_n) = 0 \) and since \( E = \bigcup_{n=1}^{\infty} E_n \), it follows \( (\mu + \lambda) (E) = 0 \). A similar argument shows that for

\[ E = \{ x \in \Omega : \text{Im} \, h(x) < 0 \}, \]

\( (\mu + \lambda)(E) = 0 \). Thus there is no loss of generality in assuming \( h \) is real-valued.

The next task is to show \( h \) is nonnegative. This is done in the same manner as above. Define the set where it is negative and then show this set has measure zero.

Let \( E \equiv \{ x : h(x) < 0 \} \) and let \( E_n \equiv \{ x : h(x) < -\frac{1}{n} \} \). Then let \( g = \mathcal{X}_{E_n} \). Since \( E = \bigcup_{n=1}^{\infty} E_n \), it follows that if \( (\mu + \lambda) (E) > 0 \) then this is also true for \( (\mu + \lambda) (E_n) \) for all \( n \) large enough. Then from 23.10

\[ \lambda(E_n) = \int_{E_n} h \, d(\mu + \lambda) \leq -(1/n) (\mu + \lambda) (E_n) < 0, \]

a contradiction. Thus it can be assumed \( h \geq 0 \). This shows that in every case,

\[ \lambda(E) = \int_E h \, d(\mu + \lambda) \]

At this point the argument splits into two cases.

**Case Where \( \lambda \ll \mu \).** In this case, \( h < 1 \).

Let \( E = [h \geq 1] \) and let \( g = \mathcal{X}_E \). Then

\[ \lambda(E) = \int_E h \, d(\mu + \lambda) \geq \mu(E) + \lambda(E). \]

Therefore \( \mu(E) = 0 \). Since \( \lambda \ll \mu \), it follows that \( \lambda(E) = 0 \) also. Thus it can be assumed

\[ 0 \leq h(x) < 1 \]

for all \( x \).

From 23.9, whenever \( g \in L^2(\Omega, \mu + \lambda) \),

\[ \int_\Omega g \, d\lambda = \int_\Omega hgd(\mu + \lambda) \]
and so
\[ \int_{\Omega} g(1 - h) d\lambda = \int_{\Omega} h g d\mu. \] (23.11)

Now let \( E \) be a measurable set and define
\[ g(x) = \sum_{i=0}^{n} h^i(x) \chi_E(x) \]
in (23.11). This yields
\[ \int_{E} (1 - h^{n+1}(x)) d\lambda = \int_{E} \sum_{i=1}^{n+1} h^i(x) d\mu. \] (23.12)

Let \( f(x) = \sum_{i=1}^{\infty} h^i(x) \) and use the Monotone Convergence theorem in (23.12) to let \( n \to \infty \) and conclude
\[ \lambda(E) = \int_{E} f \, d\mu. \]

\( f \in L^1(\Omega, \mu) \) because \( \lambda \) is finite.

The function, \( f \) is unique \( \mu \) a.e. because, if \( g \) is another function which also serves to represent \( \lambda \), consider for each \( n \in \mathbb{N} \) the set,
\[ E_n = \left\{ f - g > \frac{1}{n} \right\} \]
and conclude that
\[ 0 = \int_{E_n} (f - g) \, d\mu \geq \frac{1}{n} \mu(E_n). \]

Therefore, \( \mu(E_n) = 0 \). It follows that
\[ \mu([f - g > 0]) = \sum_{n=1}^{\infty} \mu(E_n) = 0 \]

Similarly, the set where \( g \) is larger than \( f \) has measure zero. This proves the theorem when \( \mu \gg \lambda \).

**Case where it is not necessarily true that** \( \lambda \ll \mu \).

In this case, let \( N = [h \geq 1] \) and let \( g = \chi_N \). Then
\[ \lambda(N) = \int_{N} h \, d(\mu + \lambda) \geq \mu(N) + \lambda(N). \]

and so \( \mu(N) = 0 \). Now define measures, \( \lambda_{\perp}, \lambda_{\parallel} \)
\[ \lambda_{\perp}(E) \equiv \lambda(E \cap N), \quad \lambda_{\parallel}(E) \equiv \lambda(E \cap N^C) \text{ so } \lambda = \lambda_{\perp} + \lambda_{\parallel} \]

Since \( \mu(N) = 0 \),
\[ \mu(E) = \mu(E \cap N^C) \]

Suppose then that \( \mu(E) = \mu(E \cap N^C) = 0 \). Does \( \lambda_{\parallel}(E) = 0 \)? Then since \( h < 1 \) on \( N^C \), if \( \lambda_{\parallel}(E) > 0 \),
\[ \lambda_{\parallel}(E) \equiv \lambda(E \cap N^C) = \int_{E \cap N^C} h \, d(\mu + \lambda) \]
\[ < \mu(E \cap N^C) + \lambda(E \cap N^C) = \mu(E) + \lambda_{\parallel}(E), \]

which is a contradiction because of the strict inequality which results if \( \lambda_{\parallel}(E) > 0 \). Therefore, \( \lambda_{\parallel} \ll \mu \) because if \( \mu(E) = 0 \), then \( \lambda_{\parallel}(E) = 0 \).

It only remains to verify the two measures \( \lambda_{\perp} \) and \( \lambda_{\parallel} \) are unique. Suppose then that \( \hat{\lambda}_{\perp} \) and \( \hat{\lambda}_{\parallel} \) play the roles of \( \lambda_{\perp} \) and \( \lambda_{\parallel} \) respectively. Let \( \hat{N} \) play the role of \( N \) in the definition of \( \hat{\lambda}_{\perp} \) and let \( \hat{f} \)
Suppose 23.2.2

By the monotone convergence theorem

\[ E \]

Also, for \( S \)

a nonnegative

Then both \( E \) for all

and \( S \)

Corollary 23.2.3

the Radon Nikodym derivative.

Therefore, \( E \)

and so \( E \) a.e. Let

The next corollary is a useful generalization to \( \sigma \) finite measure spaces.

Corollary 23.2.3 Suppose \( \lambda \ll \mu \) and there exist sets \( S_n \in S \), the \( \sigma \) algebra of measurable sets with

\[ S_n \cap S_m = \emptyset, \cup_{n=1}^{\infty} S_n = \Omega, \]

and \( \lambda(S_n), \mu(S_n) < \infty. \) Then there exists \( f \geq 0 \), where \( f \) is \( \mu \) measurable, and

\[ \lambda(E) = \int_E f \, d\mu \]

for all \( E \in S \). The function \( f \) is \( \mu + \lambda \) a.e. unique.

Proof: Define the \( \sigma \) algebra of subsets of \( S_n, \)

\[ S_n \equiv \{ E \cap S_n : E \in S \}. \]

Then both \( \lambda, \) and \( \mu \) are finite measures on \( S_n, \) and \( \lambda \ll \mu. \) Thus, by Theorem 23.2.2, there exists a nonnegative \( S_n \) measurable function \( f_n \), with \( \lambda(E) = \int_E f_n \, d\mu \) for all \( E \in S_n. \) Define \( f(x) = f_n(x) \) for \( x \in S_n. \) Since the \( S_n \) are disjoint and their union is all of \( \Omega, \) this defines \( f \) on all of \( \Omega. \) The function, \( f \) is measurable because

\[ f^{-1}((a, \infty]) = \cup_{n=1}^{\infty} f_n^{-1}((a, \infty]) \in S. \]

Also, for \( E \in S, \)

\[ \lambda(E) = \sum_{n=1}^{\infty} \lambda(E \cap S_n) = \sum_{n=1}^{\infty} \int \chi_{E \cap S_n}(x) f_n(x) \, d\mu \]

= \[
\sum_{n=1}^{\infty} \int \chi_{E \cap S_n}(x) f(x) \, d\mu
\]

By the monotone convergence theorem

\[
\sum_{n=1}^{\infty} \int \chi_{E \cap S_n}(x) f(x) \, d\mu = \lim_{N \to \infty} \sum_{n=1}^{N} \int \chi_{E \cap S_n}(x) f(x) \, d\mu
\]

= \[
\lim_{N \to \infty} \int \sum_{n=1}^{N} \chi_{E \cap S_n}(x) f(x) \, d\mu
\]

= \[
\int \sum_{n=1}^{\infty} \chi_{E \cap S_n}(x) f(x) \, d\mu = \int_{E} f \, d\mu.
\]
This proves the existence part of the corollary.

To see $f$ is unique, suppose $f_1$ and $f_2$ both work and consider for $n \in \mathbb{N}$

$$E_k \equiv \left[ f_1 - f_2 > \frac{1}{k} \right].$$

Then

$$0 = \lambda(E_k \cap S_n) - \lambda(E_k \cap S_n) = \int_{E_k \cap S_n} f_1(x) - f_2(x) d\mu.$$  

Hence $\mu(E_k \cap S_n) = 0$ for all $n$ so

$$\mu(E_k) = \lim_{n \to \infty} \mu(E \cap S_n) = 0.$$  

Hence $\mu([f_1 - f_2 > 0]) \leq \sum_{k=1}^{\infty} \mu(E_k) = 0$. Therefore, $\lambda([f_1 - f_2 > 0]) = 0$ also. Similarly

$$(\mu + \lambda) ([f_1 - f_2 < 0]) = 0.$$  

This version of the Radon Nikodym theorem will suffice for most applications, but more general versions are available. To see one of these, one can read the treatment in Hewitt and Stromberg [17]. This involves the notion of decomposable measure spaces, a generalization of $\sigma$-finite.

Not surprisingly, there is a simple generalization of the Lebesgue decomposition part of Theorem 23.2.2. RADON NIKODYM THEOREM

**Corollary 23.2.4** Let $(\Omega, \mathcal{S})$ be a set with a $\sigma$ algebra of sets. Suppose $\lambda$ and $\mu$ are two measures defined on the sets of $\mathcal{S}$ and suppose there exists a sequence of disjoint sets of $\mathcal{S}$, $\{\Omega_i\}_{i=1}^{\infty}$ such that $\lambda(\Omega_i), \mu(\Omega_i) < \infty, \Omega = \bigcup_i \Omega_i$. Then there is a set of $\mu$ measure zero, $N$ and measures $\lambda_\perp$ and $\lambda_{||}$ such that

$$\lambda_\perp + \lambda_{||} = \lambda, \lambda_{||} \ll \mu, \lambda_\perp (E) = \lambda(E \cap N) = \lambda_\perp (E \cap N).$$

**Proof:** Let $S_i \equiv \{E \cap \Omega_i : E \in \mathcal{S}\}$ and for $E \in S_i$, let $\lambda_i^i (E) = \lambda(E)$ and $\mu_i^i (E) = \mu(E)$. Then by Theorem 23.2.2 there exist unique measures $\lambda_\perp^i$ and $\lambda_{||}^i$ such that $\lambda_i = \lambda_\perp^i + \lambda_{||}^i$, a set of $\mu_i$ measure zero, $N_i \in S_i$ such that for all $E \in S_i$, $\lambda_i^i (E) = \lambda_i^i (E \cap N_i)$ and $\lambda_{||}^i \ll \mu_i$. Define for $E \in \mathcal{S}$

$$\lambda_\perp(E) = \sum_i \lambda_\perp^i (E \cap \Omega_i), \lambda_{||}(E) = \sum_i \lambda_{||}^i (E \cap \Omega_i), N \equiv \bigcup_i N_i.$$  

First observe that $\lambda_\perp$ and $\lambda_{||}$ are measures.

$$\lambda_\perp \left( \bigcup_{j=1}^{\infty} E_j \right) = \sum_j \lambda_\perp^i \left( \bigcup_{j=1}^{\infty} E_j \cap \Omega_i \right) = \sum_j \sum_i \lambda_\perp^i \left( E_j \cap \Omega_i \right)$$

$$= \sum_j \sum_i \lambda_\perp^i \left( E_j \cap \Omega_i \right) = \sum_j \sum_i \lambda \left( E_j \cap \Omega_i \cap N_i \right)$$

$$= \sum_j \sum_i \lambda_\perp^i \left( E_j \cap \Omega_i \right) = \sum_j \lambda_\perp \left( E_j \right).$$

The argument for $\lambda_{||}$ is similar. Now

$$\mu(N) = \sum_i \mu(N \cap \Omega_i) = \sum_i \mu_i (N_i) = 0$$

and

$$\lambda_\perp (E) = \sum_i \lambda_\perp^i (E \cap \Omega_i) = \sum_i \lambda_i (E \cap \Omega_i \cap N_i)$$

$$= \sum_i \lambda (E \cap \Omega_i \cap N) = \lambda(E \cap N).$$
Also if \( \mu(\Lambda) = 0 \), then \( \mu^i(\Lambda \cap \Omega_i) = 0 \) and so \( \lambda_\parallel^i(\Lambda \cap \Omega_i) = 0 \). Therefore,
\[
\lambda_\parallel(\Lambda) = \sum_i \lambda_\parallel^i(\Lambda \cap \Omega_i) = 0.
\]
The decomposition is unique because of the uniqueness of the \( \lambda_\parallel^i \) and \( \lambda_\parallel^j \) and the observation that some other decomposition must coincide with the given one on the \( \Omega_i \).

### 23.3 Improved Change Of Variables Formula

Recall the change of variables formula. An assumption was made that the transformation was \( C^1 \). This made the argument easy to push through with the use of the Besicovitch covering theorem. However, with the Radon Nikodym theorem and the fundamental theorem of calculus, it is easy to give a much shorter argument which actually gives a better result because it does not require the derivative to be continuous. Recall that \( U \subseteq \mathbb{R}^n \), was an open set on which \( h \) was differentiable and one to one with \( \det(Dh(x)) \neq 0 \). Then for \( x \in U \), we had an inequality which came from the definition of the derivative and the Brouwer fixed point theorem.

\[
|\det Dh(x)| (1 - \varepsilon)^n \leq \frac{m_n(h(B(x,r_x)))}{m_n(B(x,r_x))} \leq |\det Dh(x)| (1 + \varepsilon)^n
\]

this for all \( r_x \) sufficiently small. Now consider the following measure \( \mu(\Lambda) \equiv m_n(h(\Lambda)) \). By Proposition 21.5.1 this is well defined. It is indeed a measure and \( \mu \ll m_n \) the details in Problem 21 on Page 533. Therefore, by Corollary 23.2.3, there is a nonnegative measurable \( g \) which is in \( L^1_{\text{loc}}(\mathbb{R}^n) \) such that

\[
\mu(\Lambda) = \int_{\Lambda} g \, dm_n
\]

It is clearly a Radon measure. In particular,

\[
m_n(h(B(x,r))) = \int_{B(x,r)} g \, dm_n
\]

The idea is to identify \( g \). Let \( x \) be a Lebesgue point of \( g \), for all \( r \) small enough,

\[
|\det Dh(x)| (1 - \varepsilon)^n \leq \frac{m_n(h(B(x,r)))}{m_n(B(x,r))} = \frac{1}{m_n(B(x,r))} \int_{B(x,r)} g \, dm_n \leq |\det Dh(x)| (1 + \varepsilon)^n
\]

Then letting \( r \to 0 \),

\[
|\det Dh(x)| (1 - \varepsilon)^n \leq g(x) \leq |\det Dh(x)| (1 + \varepsilon)^n
\]

and so, since \( \varepsilon \) is arbitrary, \( g(x) = |\det Dh(x)| \). Thus for any measurable \( A \subseteq U \),

\[
m_n(h(A)) = \int_A |\det Dh| \, dm_n \tag{**}
\]

This yields the following theorem.

**Theorem 23.3.1** Let \( U \) be an open set in \( \mathbb{R}^n \) and let \( h : U \to \mathbb{R}^n \) be one to one and differentiable with \( \det Dh(x) \neq 0 \). Then if \( f \geq 0 \) and Lebesgue measurable,

\[
\int_{h(U)} f \, dm_n = \int_U (f \circ h) |\det Dh| \, dm_n
\]
Proof: The proof is exactly as before. You just approximate with simple functions and use **. Now if \( A \) is measurable, so is \( h(A) \) by Proposition 21.5.1. Then from the above,

\[
\int_{h(U)} X_{h(A)}(y) f(y) \, dm = \int_U (X_A(x)(f \circ h)(x)) |\det Dh(x)| \, dm_n
\]

Thus

\[
\int_{h(A)} f(y) \, dm = \int_A (f \circ h)(x) |\det Dh(x)| \, dm_n
\]

This gives a slight generalization.

Corollary 23.3.2 Let \( A \) be any measurable subset of \( U \) and let \( f \) be a nonnegative Lebesgue measurable function. Let \( h : U \to \mathbb{R}^n \) be one to one and differentiable with \( \det Dh(x) \neq 0 \). Then

\[
\int_{h(A)} f(y) \, dm = \int_A (f \circ h)(x) |\det Dh(x)| \, dm_n
\]

As to the generalization when \( h \) is only assumed to be one to one and differentiable, this also follows as before. The proof of Sard’s lemma in Theorem 21.5.6 did not require any continuity of the derivative. Thus as before, \( h(U_0) = 0 \) where \( U_0 \equiv \{ x : \det Dh(x) = 0 \} \), Borel measurable because \( \det Dh \) is Borel measurable due to the fact that \( h \) is continuous and the entries of the matrix of \( Dh \) are each Borel measurable because they are limits of difference quotients which are continuous functions. Letting \( U_+ = \{ x : |\det Dh(x)| > 0 \} \), Corollary 23.3.2 and Sard’s lemma imply

\[
\int_{h(U)} f(y) \, dm = \int_{h(U_+)} f(y) \, dm = \int_{U_+} (f \circ h)(x) |\det Dh(x)| \, dm_n = \int_U (f \circ h)(x) |\det Dh(x)| \, dm_n
\]

Then one can also obtain the result of Corollary 23.3.2 in the same way. This leads to the following version of the change of variables formula.

Theorem 23.3.3 Let \( A \) be any measurable subset of \( U \), an open set in \( \mathbb{R}^n \) and let \( f \) be a nonnegative Lebesgue measurable function. Let \( h : U \to \mathbb{R}^n \) be one to one and differentiable. Then

\[
\int_{h(A)} f(y) \, dm = \int_A (f \circ h)(x) |\det Dh(x)| \, dm_n
\]

You can also tweak this a little more to get a slightly more general result. You could assume, for example, that \( h \) is continuous on the open set \( U \) and differentiable and one to one on some \( F \subseteq U \) where \( F \) is measurable and \( m_n(h(U \setminus F)) = 0 \) rather than assuming that \( h \) is differentiable on all of the open set \( U \). This would end up working out also, but the above is pretty close and is easier to remember.

23.4 Vector Measures

The next topic will use the Radon Nikodym theorem. It is the topic of vector and complex measures. The main interest is in complex measures although a vector measure can have values in any topological vector space. Whole books have been written on this subject. See for example the book by Diestal and Uhl titled Vector measures.
Definition 23.4.1 Let \((V, \| \cdot \|)\) be a normed linear space and let \((\Omega, S)\) be a measure space. A function \(\mu : S \to V\) is a vector measure if \(\mu\) is countably additive. That is, if \(\{E_i\}_{i=1}^\infty\) is a sequence of disjoint sets of \(S\),

\[
\mu(\bigcup_{i=1}^\infty E_i) = \sum_{i=1}^\infty \mu(E_i).
\]

Note that it makes sense to take finite sums because it is given that \(\mu\) has values in a vector space in which vectors can be summed. In the above, \(\mu(E_i)\) is a vector. It might be a point in \(\mathbb{R}^n\) or in any other vector space. In many of the most important applications, it is a vector in some sort of function space which may be infinite dimensional. The infinite sum has the usual meaning. That is

\[
\sum_{i=1}^\infty \mu(E_i) = \lim_{n \to \infty} \sum_{i=1}^n \mu(E_i)
\]

where the limit takes place relative to the norm on \(V\).

Definition 23.4.2 Let \((\Omega, S)\) be a measure space and let \(\mu\) be a vector measure defined on \(S\). A subset, \(\pi(E)\), of \(S\) is called a partition of \(E\) if \(\pi(E)\) consists of finitely many disjoint sets of \(S\) and \(\cup \pi(E) = E\). Let

\[
|\mu|(E) = \sup\{ \sum_{F \in \pi(E)} \|\mu(F)\| : \pi(E) \text{ is a partition of } E \}.
\]

|\mu| is called the total variation of \(\mu\).

The next theorem may seem a little surprising. It states that, if finite, the total variation is a nonnegative measure.

Theorem 23.4.3 If \(|\mu|(\Omega) < \infty\), then \(|\mu|\) is a measure on \(S\). Even if \(|\mu|(\Omega) = \infty\), \(|\mu|(\bigcup_{i=1}^\infty E_i) \leq \sum_{i=1}^\infty |\mu|(E_i)\). That is \(|\mu|\) is subadditive and \(|\mu|(A) \leq |\mu|(B)\) whenever \(A, B \in S\) with \(A \subseteq B\).

Proof: Consider the last claim. Let \(a < |\mu|(A)\) and let \(\pi(A)\) be a partition of \(A\) such that

\[
a < \sum_{F \in \pi(A)} \|\mu(F)\|.
\]

Then \(\pi(A) \cup \{B \setminus A\}\) is a partition of \(B\) and

\[
|\mu|(B) \geq \sum_{F \in \pi(A)} \|\mu(F)\| + \|\mu(B \setminus A)\| > a.
\]

Since this is true for all such \(a\), it follows \(|\mu|(B) \geq |\mu|(A)\) as claimed.

Let \(\{E_j\}_{j=1}^\infty\) be a sequence of disjoint sets of \(S\) and let \(E_\infty = \bigcup_{j=1}^\infty E_j\). Then letting \(a < |\mu|(E_\infty)\), it follows from the definition of total variation there exists a partition of \(E_\infty\), \(\pi(E_\infty) = \{A_1, \cdots, A_n\}\) such that

\[
a < \sum_{i=1}^n \|\mu(A_i)\|.
\]

Also,

\[
A_i = \bigcup_{j=1}^\infty A_i \cap E_j
\]

and so by the triangle inequality, \(\|\mu(A_i)\| \leq \sum_{j=1}^\infty \|\mu(A_i \cap E_j)\|\). Therefore, by the above, and either Fubini’s theorem or Lemma 23.3 on Page 566.

\[
a < \sum_{i=1}^n \sum_{j=1}^\infty \|\mu(A_i \cap E_j)\| = \sum_{j=1}^\infty \sum_{i=1}^n \|\mu(A_i \cap E_j)\| \leq \sum_{j=1}^\infty |\mu|(E_j)
\]
23.4. VECTOR MEASURES

because \( \{ A_i \cap E_j \}_{i=1}^n \) is a partition of \( E_j \).

Since \( a \) is arbitrary, this shows

\[
|\mu|((\bigcup_{j=1}^\infty E_j) \leq \sum_{j=1}^\infty |\mu|(E_j).
\]

If the sets, \( E_j \) are not disjoint, let \( F_1 = E_1 \) and if \( F_n \) has been chosen, let \( F_{n+1} = E_{n+1} \setminus \bigcup_{i=1}^n E_i \). Thus the sets, \( F_i \) are disjoint and \( \bigcup_{i=1}^\infty F_i = \bigcup_{i=1}^\infty E_i \). Therefore,

\[
|\mu| \left( \bigcup_{j=1}^\infty E_j \right) = |\mu| \left( \bigcup_{j=1}^\infty F_j \right) \leq \sum_{j=1}^\infty |\mu|(F_j) \leq \sum_{j=1}^\infty |\mu|(E_j)
\]

and proves \( |\mu| \) is always subadditive as claimed regardless of whether \( |\mu|(\Omega) < \infty \).

Now suppose \( |\mu|(\Omega) < \infty \) and let \( E_1 \) and \( E_2 \) be sets of \( S \) such that \( E_1 \cap E_2 = \emptyset \) and let \( \{ A_1 \ldots A_n \} = \pi(E_i) \), a partition of \( E_i \) which is chosen such that

\[
|\mu|(E_i) - \varepsilon < \sum_{j=1}^n |\mu(A_j)|| i = 1, 2.
\]

Such a partition exists because of the definition of the total variation and \( |\mu|(\Omega) < \infty \). Consider the sets which are contained in either of \( \pi(E_1) \) or \( \pi(E_2) \), it follows this collection of sets is a partition of \( E_1 \cup E_2 \) denoted by \( \pi(E_1 \cup E_2) \). Then by the above inequality and the definition of total variation,

\[
|\mu|(E_1 \cup E_2) \geq \sum_{F \in \pi(E_1 \cup E_2)} |\mu(F)|| > |\mu|(E_1) + |\mu|(E_2) - 2\varepsilon,
\]

which shows that since \( \varepsilon > 0 \) was arbitrary,

\[
|\mu|(E_1 \cup E_2) \geq |\mu|(E_1) + |\mu|(E_2). \quad (23.13)
\]

Then implies that whenever the \( E_i \) are disjoint, \( |\mu|(\bigcup_{j=1}^n E_j) \geq \sum_{j=1}^n |\mu|(E_j) \). Therefore, for the \( E_i \) disjoint,

\[
\sum_{j=1}^\infty |\mu|(E_j) \geq |\mu|(\bigcup_{j=1}^\infty E_j) \geq |\mu|(\bigcup_{j=1}^n E_j) \geq \sum_{j=1}^n |\mu|(E_j).
\]

Since \( n \) is arbitrary,

\[
|\mu|(\bigcup_{j=1}^\infty E_j) = \sum_{j=1}^\infty |\mu|(E_j)
\]

which shows that \( |\mu| \) is a measure as claimed. \( \blacksquare \)

The following corollary is interesting. It concerns the case that \( \mu \) is only finitely additive.

**Corollary 23.4.4** Suppose \( (\Omega, \mathcal{F}) \) is a set with a \( \sigma \) algebra of subsets \( \mathcal{F} \) and suppose \( \mu : \mathcal{F} \to \mathbb{C} \) is only finitely additive. That is, \( \mu(\bigcup_{i=1}^n E_i) = \sum_{i=1}^n \mu(E_i) \) whenever the \( E_i \) are disjoint. Then \( |\mu| \), defined in the same way as above, is also finitely additive provided \( |\mu| \) is finite.

**Proof:** Say \( E \cap F = \emptyset \) for \( E, F \in \mathcal{F} \). Let \( \pi(E), \pi(F) \) suitable partitions for which the following holds.

\[
|\mu|(E \cup F) \geq \sum_{A \in \pi(E)} |\mu(A)| + \sum_{B \in \pi(F)} |\mu(B)| \geq |\mu|(E) + |\mu|(F) - 2\varepsilon.
\]

Since \( \varepsilon \) is arbitrary, \( |\mu|(E \cap F) \geq |\mu|(E) + |\mu|(F) \). Similar considerations apply to any finite union of disjoint sets. That is, if the \( E_i \) are disjoint, then

\[
|\mu|(\bigcup_{i=1}^n E_i) \geq \sum_{i=1}^n |\mu|(E_i).
\]
Now let $E = \bigcup_{i=1}^{n} E_i$ where the $E_i$ are disjoint. Then letting $\pi (E)$ be a partition of $E$,

$$|\mu| (E) - \varepsilon \leq \sum_{F \in \pi (E)} |\mu (F)| ,$$

it follows that

$$|\mu| (E) \leq \varepsilon + \sum_{F \in \pi (E)} |\mu (F)| = \varepsilon + \sum_{i=1}^{n} \sum_{F \in \pi (E)} |\mu (F \cap E_i)| .$$

$$\leq \varepsilon + \sum_{i=1}^{n} \sum_{F \in \pi (E)} |\mu (F \cap E_i)| \leq \varepsilon + \sum_{i=1}^{n} |\mu (E_i)| .$$

Since $\varepsilon$ is arbitrary, this shows $|\mu| (\bigcup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} |\mu (E_i)|$. Thus $|\mu|$ is finitely additive.

**Theorem 23.4.5** Suppose $\mu$ is a complex measure on $(\Omega, \mathcal{S})$ where $\mathcal{S}$ is a $\sigma$ algebra of subsets of $\Omega$. That is, whenever, $\{E_i\}$ is a sequence of disjoint sets of $\mathcal{S}$,

$$\mu (\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu (E_i) .$$

Then $|\mu| (\Omega) < \infty$.

**Proof:** First here is a claim.

**Claim:** Suppose $|\mu| (E) = \infty$. Then there are disjoint subsets of $E$, $A$ and $B$ such that $E = A \cup B$, $|\mu (A)|, |\mu (B)| > 1$ and $|\mu| (B) = \infty$.

**Proof of the claim:** From the definition of $|\mu|$, there exists a partition of $E, \pi (E)$ such that

$$\sum_{F \in \pi (E)} |\mu (F)| > 20 (1 + |\mu (E)|) .$$

(23.14)

Here 20 is just a nice sized number. No effort is made to be delicate in this argument. Also note that $\mu (E) \in C$ because it is given that $\mu$ is a complex measure. Consider the following picture consisting of two lines in the complex plane having slopes 1 and $-1$ which intersect at the origin, dividing the complex plane into four closed sets, $R_1, R_2, R_3$, and $R_4$ as shown. Let $\pi_i$ consist of those sets, $A$ of $\pi (E)$ for which $\mu (A) \in R_i$. Thus, some sets, $A$ of $\pi (E)$ could be in two of the $\pi_i$ if $\mu (A)$ is one of the intersecting lines. This is not important. The thing which is important is that if $\mu (A) \in R_1$ or $R_3$, then $\sqrt{2} |\mu (A)| \leq |\text{Re} (\mu (A))|$ and if $\mu (A) \in R_2$ or $R_4$ then $\frac{\sqrt{2}}{2} |\mu (A)| \leq |\text{Im} (\mu (A))|$ and both $R_1$ and $R_3$ have complex numbers $z$ contained in these sets all have the same sign for $\text{Re} (z)$. Thus, for $z_i \in R_1$, $\sum_i \text{Re} (z_i) = \sum_i |\text{Re} (z_i)|$. A similar statement holds for $z_i \in R_3$. In the case of $R_2, R_4$, similar considerations hold for the imaginary parts. Thus $|\sum_i \text{Im} z_i| = |\sum_i |\text{Im} z_i| |$ is $z_i$ are all in $R_2$ or else all in $R_4$. Then by (23.14), it follows that for some $i$,

$$\sum_{F \in \pi_i} |\mu (F)| > 5 (1 + |\mu (E)|) .$$

(23.15)

Suppose $i$ equals 1 or 3. A similar argument using the imaginary part applies if $i$ equals 2 or 4. Then, since $\text{Re} (\mu (F))$ always has the same sign,

$$\left| \sum_{F \in \pi_i} \mu (F) \right| \geq \left| \sum_{F \in \pi_i} \text{Re} (\mu (F)) \right| = \sum_{F \in \pi_i} |\text{Re} (\mu (F))| \geq \frac{\sqrt{2}}{2} \sum_{F \in \pi_i} |\mu (F)| > 5 \frac{\sqrt{2}}{2} (1 + |\mu (E)|) .$$
Now let \( C \) be the union of the sets in \( \pi_i \),
\[
|\mu(C)| = \left| \sum_{F \in \pi_i} \mu(F) \right| > \frac{5}{2} (1 + |\mu(E)|) > 1. \tag{23.16}
\]
Define \( D \) such that \( C \cup D = E \). Then, since \( \mu \) is a measure,
\[
\frac{5}{2} (1 + |\mu(E)|) < |\mu(C)| = |\mu(E) - \mu(D)|
\leq |\mu(E)| + |\mu(D)|
\]
and so, subtracting \( |\mu(E)| \) from both sides,
\[
1 < \frac{5}{2} + \frac{3}{2} |\mu(E)| < |\mu(D)|.
\]
Now since \( |\mu|(\Omega) = \infty \), it follows from Theorem 23.4.6 that \( \infty = |\mu|(\Omega) \leq |\mu(C)| + |\mu(D)| \) and so either \( |\mu(C)| = \infty \) or \( |\mu(D)| = \infty \). If \( |\mu(C)| = \infty \), let \( B = C \) and \( A = D \). Otherwise, let \( B = D \) and \( A = C \). This proves the claim.

Now suppose \( |\mu|(\Omega) = \infty \). Then from the claim, there exist \( A_1 \) and \( B_1 \) such that \( |\mu(B_1)| = \infty, |\mu(B_1)|, |\mu(A_1)| > 1, \) and \( A_1 \cup B_1 = \Omega \). Let \( B_1 \equiv \Omega \setminus A \) play the same role as \( \Omega \) and obtain \( A_2, B_2 \subseteq B_1 \) such that \( |\mu(B_2)| = \infty, |\mu(B_2)|, |\mu(A_2)| > 1, \) and \( A_2 \cup B_2 = B_1 \). Continue in this way to obtain a sequence of disjoint sets, \( \{A_i\} \) such that \( |\mu(A_i)| > 1 \). Then since \( \mu \) is a measure,
\[
\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)
\]
but this is impossible because \( \lim_{i \to \infty} \mu(A_i) \neq 0 \).

**Theorem 23.4.6** Let \( (\Omega, S) \) be a measure space and let \( \lambda : S \to \mathbb{C} \) be a complex vector measure. Thus \( |\lambda|(\Omega) < \infty \). Let \( \mu : S \to [0, \mu(\Omega)] \) be a finite measure such that \( \lambda \ll \mu \). Then there exists a unique \( f \in L^1(\Omega) \) such that for all \( E \in S \),
\[
\int_E f \, d\mu = \lambda(E).
\]

**Proof:** It is clear that \( \text{Re} \lambda \) and \( \text{Im} \lambda \) are real-valued vector measures on \( S \). Since \( |\lambda|(\Omega) < \infty \), it follows easily that \( |\text{Re} \lambda|(\Omega) \) and \( |\text{Im} \lambda|(\Omega) < \infty \). This is clear because
\[
|\lambda(E)| \geq |\text{Re} \lambda(E)|, |\text{Im} \lambda(E)|.
\]
Therefore, each of
\[
\frac{|\text{Re} \lambda| + \text{Re} \lambda}{2}, \quad \frac{|\text{Re} \lambda| - \text{Re} \lambda}{2}, \quad \frac{|\text{Im} \lambda| + \text{Im} \lambda}{2}, \quad \text{and} \quad \frac{|\text{Im} \lambda| - \text{Im} \lambda}{2}
\]
are finite measures on \( S \). It is also clear that each of these finite measures are absolutely continuous with respect to \( \mu \) and so there exist unique nonnegative functions in \( L^1(\Omega) \), \( f_1, f_2, g_1, g_2 \) such that for all \( E \in S \),
\[
\frac{1}{2}(|\text{Re} \lambda| + \text{Re} \lambda)(E) = \int_E f_1 \, d\mu,
\]
\[
\frac{1}{2}(|\text{Re} \lambda| - \text{Re} \lambda)(E) = \int_E f_2 \, d\mu,
\]
\[
\frac{1}{2}(|\text{Im} \lambda| + \text{Im} \lambda)(E) = \int_E g_1 \, d\mu,
\]
\[
\frac{1}{2}(|\text{Im} \lambda| - \text{Im} \lambda)(E) = \int_E g_2 \, d\mu.
\]
Now let \( f = f_1 - f_2 + i(g_1 - g_2) \).

The following corollary is about representing a vector measure in terms of its total variation. It is like representing a complex number in the form \( re^{i\theta} \). The proof requires the following lemma.
Lemma 23.4.7 Suppose \((\Omega, S, \mu)\) is a measure space and \(f\) is a function in \(L^1(\Omega, \mu)\) with the property that
\[
|\int_E f \, d\mu| \leq \mu(E)
\]
for all \(E \in S\). Then \(|f| \leq 1 \text{ a.e.}\)

Proof of the lemma: Consider the following picture where \(B(p, r) \cap B(0, 1) \neq \emptyset\). Let \(E = f^{-1}(B(p, r))\). In fact, \(\mu(E) = 0\). If \(\mu(E) \neq 0\) then
\[
\frac{1}{\mu(E)} \int_E f \, d\mu = \frac{1}{\mu(E)} \int_E (f-p) \, d\mu \leq \frac{1}{\mu(E)} \int_E |f-p| \, d\mu < r
\]
because on \(E\), \(|f(\omega) - p| < r\). Hence \(\frac{1}{\mu(E)} \int_E f \, d\mu\) is closer to \(p\) than \(r\) and so
\[
\frac{1}{\mu(E)} \int_E f \, d\mu > 1
\]
Refer to the picture. However, this contradicts the assumption of the lemma. It follows \(\mu(E) = 0\).

Since the set of complex numbers, \(z\) such that \(|z| > 1\) is an open set, it equals the union of countably many balls, \(\{B_i\}_{i=1}^\infty\). Therefore,
\[
\mu\left(f^{-1}\{\{z \in \mathbb{C} : |z| > 1\}\}\right) = \mu\left(\bigcup_{k=1}^\infty f^{-1}(B_k)\right) \\
\leq \sum_{k=1}^\infty \mu\left(f^{-1}(B_k)\right) = 0.
\]
Thus \(|f(\omega)| \leq 1 \text{ a.e. as claimed.}\)

Corollary 23.4.8 Let \(\lambda\) be a complex vector measure with \(|\lambda|(\Omega) < \infty\). Then there exists a unique \(f \in L^1(\Omega)\) such that \(\lambda(E) = \int_E f \, d|\lambda|\). Furthermore, \(|f| = 1\) for \(|\lambda| \text{ a.e.}\) This is called the polar decomposition of \(\lambda\).

Proof: First note that \(\lambda \ll |\lambda|\) and so such an \(L^1\) function exists and is unique. It is required to show \(|f| = 1 \text{ a.e.}\) If \(|\lambda|(E) \neq 0\),
\[
\left|\frac{\lambda(E)}{|\lambda|(E)}\right| = \frac{1}{|\lambda|(E)} \int_E f \, d|\lambda| \leq 1.
\]
Therefore by Lemma 23.4.7, \(|f| \leq 1, |\lambda| \text{ a.e.}\) Now let
\[
E_n = \left[|f| \leq 1 - \frac{1}{n}\right].
\]
Let \(\{F_1, \cdots, F_m\}\) be a partition of \(E_n\). Then
\[
\sum_{i=1}^m |\lambda(F_i)| = \sum_{i=1}^m \int_{F_i} f \, d|\lambda| \leq \sum_{i=1}^m \int_{F_i} |f| \, d|\lambda| \\
\leq \sum_{i=1}^m \int_{F_i} \left(1 - \frac{1}{n}\right) \, d|\lambda| = \sum_{i=1}^m \left(1 - \frac{1}{n}\right) |\lambda|(F_i) \\
= |\lambda|(E_n) \left(1 - \frac{1}{n}\right).
\]
\footnote{As proved above, the assumption that \(|\lambda|(\Omega) < \infty\) is redundant.}
Then taking the supremum over all partitions,
\[ |\lambda|(E_n) \leq \left(1 - \frac{1}{n}\right) |\lambda|(E_n) \]
which shows \( |\lambda|(E_n) = 0 \). Hence \( |\lambda|(\{f < 1\}) = 0 \) because \( \bigcup_{n=1}^{\infty} E_n \).

If \( \lambda(E) = \int_E \mu d\mu \), \( \mu \) a measure, it is also true that \( \lambda(E) = \int_E gd|\lambda| \). How do \( h \) and \( g \) compare?

**Corollary 23.4.9** Suppose \((\Omega, S)\) is a measure space and \( \mu \) is a finite nonnegative measure on \( S \). Then for \( h \in L^1(\mu) \), define a complex measure \( \lambda \) by
\[ \lambda(E) = \int_E h d\mu. \]
Then
\[ |\lambda|(E) = \int_E |h| d\mu. \]
Furthermore, \( |h| = \overline{g}h \) where \( gd|\lambda| \) is the polar decomposition of \( \lambda \),
\[ \lambda(E) = \int_E gd|\lambda| \]

**Proof:** From Corollary 23.4.8 there exists \( g \) such that \( |g| = 1, |\lambda| \) a.e. and for all \( E \in S \)
\[ \lambda(E) = \int_E gd|\lambda| = \int_E h d\mu. \]
Let \( s_n \) be a sequence of simple functions converging pointwise to \( \overline{g}, |s_n| \leq 1 \). Then from the above,
\[ \int_E g s_n d|\lambda| = \int_E s_n h d\mu. \]
Passing to the limit using the dominated convergence theorem,
\[ |\lambda|(E) = \int_E d|\lambda| = \int_E \overline{g}h d\mu. \]
It follows from the same kind of arguments given above that \( \overline{g}h \geq 0 \) a.e. and \( |\overline{g}| = 1 \). Therefore, \( |h| = |\overline{g}h| = \overline{g}h \). It follows from the above, that
\[ |\lambda|(E) = \int_E d|\lambda| = \int_E \overline{g}h d\mu = \int_E |h| d\mu. \]

---

### 23.5 Representation Theorems For The Dual Space Of \( L^p \)

The next topic deals with the dual space of \( L^p \) for \( p \geq 1 \) in the case where the measure space is \( \sigma \) finite or finite. In what follows \( q = \infty \) if \( p = 1 \) and otherwise, \( \frac{1}{p} + \frac{1}{q} = 1 \). Recall that the dual space of \( X \) is \( \mathcal{L}(X, \mathbb{C}) \). In fact, this theorem holds without the assumption that the measure space is \( \sigma \) finite in case \( p > 1 \) but this is more technical to establish.

**Theorem 23.5.1** (Riesz representation theorem) Let \( p > 1 \) and let \( (\Omega, S, \mu) \) be a finite measure space. If \( \Lambda \in (L^p(\Omega))' \), then there exists a unique \( h \in L^q(\Omega) \) \( \left(\frac{1}{p} + \frac{1}{q} = 1\right) \) such that
\[ \Lambda = \int_{\Omega} hf d\mu. \]
This function satisfies \( ||h||_q = ||\Lambda|| \) where \( ||\Lambda|| \) is the operator norm of \( \Lambda \).
Proof: (Uniqueness) If \( h_1 \) and \( h_2 \) both represent \( \Lambda \), consider
\[
f = |h_1 - h_2|^q (\overline{h_1} - \overline{h_2}),
\]
where \( \overline{h} \) denotes complex conjugation. By Hölder’s inequality, it is easy to see that \( f \in L^p(\Omega) \). Thus
\[
0 = \Lambda f - \Lambda f = \int h_1|h_1 - h_2|^q (\overline{h_1} - \overline{h_2}) - h_2|h_1 - h_2|^q (\overline{h_1} - \overline{h_2}) d\mu
\]
\[
= \int |h_1 - h_2|^q d\mu.
\]
Therefore \( h_1 = h_2 \) and this proves uniqueness.

Now let \( \lambda(E) = \Lambda(\mathcal{X}_E) \). Since this is a finite measure space \( \mathcal{X}_E \) is an element of \( L^p(\Omega) \) and so it makes sense to write \( \Lambda(\mathcal{X}_E) \). In fact \( \lambda \) is a complex measure having finite total variation. This follows from an earlier result but I will show it directly here. Let \( A_1, \ldots, A_n \) be a partition of \( \Omega \):
\[
|\Lambda \mathcal{X}_{A_i}| = w_i(\Lambda \mathcal{X}_{A_i}) = \Lambda(w_i \mathcal{X}_{A_i})
\]
for some \( w_i \in \mathbb{C}, \ |w_i| = 1 \). Thus
\[
\sum_{i=1}^n |\lambda(A_i)| = \sum_{i=1}^n |\Lambda(\mathcal{X}_{A_i})| = \Lambda(\sum_{i=1}^n w_i \mathcal{X}_{A_i})
\]
\[
\leq ||\lambda|| (\int \sum_{i=1}^n w_i \mathcal{X}_{A_i} |^p d\mu)^{\frac{1}{p}} = ||\lambda|| (\int \mu(\Omega) d\mu)^{\frac{1}{p}}.
\]
This is because if \( x \in \Omega, \ x \) is contained in exactly one of the \( A_i \) and so the absolute value of the sum in the first integral above is equal to 1. Therefore \( ||\lambda|| (\Omega) \leq \infty \) because this was an arbitrary partition. Also, if \( \{E_i\}_{i=1}^{\infty} \) is a sequence of disjoint sets of \( S \), let
\[
F_n = \bigcup_{i=1}^n E_i, \ F = \bigcup_{i=1}^{\infty} E_i.
\]
Then by the Dominated Convergence theorem,
\[
||\Lambda F_n - \Lambda F||_p \to 0.
\]
Therefore, by continuity of \( \Lambda \),
\[
\lambda(F) = \Lambda(\mathcal{X}_F) = \lim_{n \to \infty} \Lambda(\mathcal{X}_{F_n}) = \lim_{n \to \infty} \sum_{k=1}^n \Lambda(\mathcal{X}_{E_k}) = \sum_{k=1}^{\infty} \lambda(E_k).
\]
This shows \( \lambda \) is a complex measure with \( |\lambda| \) finite.

It is also clear from the definition of \( \lambda \) that \( \lambda \ll \mu \). Therefore, by the Radon Nikodym theorem, there exists \( h \in L^1(\Omega) \) with
\[
\lambda(E) = \int_E h d\mu = \Lambda(\mathcal{X}_E).
\]
Actually \( h \in L^q \) and satisfies the other conditions above. Let \( s = \sum_{i=1}^m c_i \mathcal{X}_{e_i} \) be a simple function. Then since \( \Lambda \) is linear,
\[
\Lambda(s) = \sum_{i=1}^m c_i \Lambda(\mathcal{X}_{e_i}) = \sum_{i=1}^m c_i \int_{E_i} h d\mu = \int h s d\mu. \quad (23.17)
\]

Claim: If \( f \) is uniformly bounded and measurable, then
\[
\Lambda(f) = \int h f d\mu.
\]
23.5. REPRESENTATION THEOREMS FOR THE DUAL SPACE OF $L^P$

**Proof of claim:** Since $f$ is bounded and measurable, there exists a sequence of simple functions, \( \{s_n\} \) which converges to $f$ pointwise and in $L^p(\Omega)$. This follows from Theorem 23.1.4 on Page 464 upon breaking $f$ up into positive and negative parts of real and complex parts. In fact this theorem gives uniform convergence. Then

$$\Lambda(f) = \lim_{n \to \infty} \Lambda(s_n) = \lim_{n \to \infty} \int hs_n d\mu = \int hf d\mu,$$

the first equality holding because of continuity of $\Lambda$, the second following from 23.17 and the third holding by the dominated convergence theorem.

This is a very nice formula but it still has not been shown that $h \in L^q(\Omega)$.

Let $E_n = \{x : |h(x)| \leq n\}$. Thus $|h\chi_{E_n}| \leq n$. Then

$$|h\chi_{E_n}|^{q-2}(\chi_{E_n}) \in L^p(\Omega).$$

By the claim, it follows that

$$||h\chi_{E_n}||_{q}^q = \int |h\chi_{E_n}|^{q-2}(\chi_{E_n}) d\mu = \Lambda(|h\chi_{E_n}|^{q-2}(\chi_{E_n}))$$

$$\leq ||\Lambda|| \ |||h\chi_{E_n}|^{q-2}(\chi_{E_n})||_p = ||\Lambda|| \ ||h\chi_{E_n}||_{q}^q,$$

the last equality holding because $q - 1 = q/p$ and so

$$\left(\int |h\chi_{E_n}|^{q-2}(\chi_{E_n})^p d\mu\right)^{1/p} = \left(\int (|h\chi_{E_n}|^{q/p})^p d\mu\right)^{1/p} = ||h\chi_{E_n}||_{q}^q.$$

Therefore, since $q - \frac{q}{p} = 1$, it follows that

$$||h\chi_{E_n}||_q \leq ||\Lambda||.$$

Letting $n \to \infty$, the Monotone Convergence theorem implies

$$||h||_q \leq ||\Lambda||. \quad (23.18)$$

Now that $h$ has been shown to be in $L^q(\Omega)$, it follows from 23.17 and the density of the simple functions, Theorem 22.2.1 on Page 541, that

$$\Lambda = \int hf d\mu$$

for all $f \in L^p(\Omega)$.

It only remains to verify the last claim.

$$||\Lambda|| = \sup\{\int hf : ||f||_p \leq 1\} \leq ||h||_q \leq ||\Lambda||$$

by 23.18 and holder’s inequality. ■

To represent elements of the dual space of $L^1(\Omega)$, another Banach space is needed.

**Definition 23.5.2** Let $(\Omega, \mathcal{S}, \mu)$ be a measure space. $L^\infty(\Omega)$ is the vector space of measurable functions such that for some $M > 0$, $|f(x)| \leq M$ for all $x$ outside of some set of measure zero ($|f(x)| \leq M$ a.e.). Define $f = g$ when $f(x) = g(x)$ a.e. and $||f||_\infty \equiv \inf\{M : |f(x)| \leq M \ a.e.\}$.

**Theorem 23.5.3** $L^\infty(\Omega)$ is a Banach space.
Theorem 23.5.4 (Riesz representation theorem) Let \((\Omega, S, \mu)\) be a finite measure space. If \(\Lambda \in (L^1(\Omega))'\), then there exists a unique \(h \in L^\infty(\Omega)\) such that

\[
\Lambda(f) = \int_\Omega hf \, d\mu
\]

for all \(f \in L^1(\Omega)\). If \(h\) is the function in \(L^\infty(\Omega)\) representing \(\Lambda \in (L^1(\Omega))'\), then \(\|h\|_\infty = ||\Lambda||\).

Proof: Just as in the proof of Theorem 23.5.1, there exists a unique \(h \in L^1(\Omega)\) such that for all simple functions \(s\),

\[
\Lambda(s) = \int hs \, d\mu. \tag{23.19}
\]
23.5. REPRESENTATION THEOREMS FOR THE DUAL SPACE OF $L^p$

To show $h \in L^\infty(\Omega)$, let $\varepsilon > 0$ be given and let

$$E = \{x : |h(x)| \geq \|\Lambda\| + \varepsilon\}.$$  

Let $|k| = 1$ and $hk = |h|$. Since the measure space is finite, $k \in L^1(\Omega)$. As in Theorem 23.5.1 let \( \{s_n\} \) be a sequence of simple functions converging to $k$ in $L^1(\Omega)$, and pointwise. It follows from the construction in Theorem 19.1.6 on Page 464 that it can be assumed $|s_n| \leq 1$. Therefore

$$\Lambda(k_XE) = \lim_{n \to \infty} \Lambda(s_nX_E) = \lim_{n \to \infty} \int_E h s_n d\mu = \int_E h k d\mu$$

where the last equality holds by the Dominated Convergence theorem. Therefore,

$$\|\Lambda\|_\mu(E) \geq |\Lambda(k_XE)| = |\int \Omega h k X_E d\mu| = \int_E |h| d\mu \geq (\|\Lambda\| + \varepsilon)\mu(E).$$

It follows that $\mu(E) = 0$. Since $\varepsilon > 0$ was arbitrary, $\|\Lambda\| \geq \|h\|_\infty$. Since $h \in L^\infty(\Omega)$, the density of the simple functions in $L^1(\Omega)$ and 23.19 imply

$$\Lambda f = \int_\Omega h f d\mu, \|\Lambda\| \geq \|h\|_\infty.$$  \hspace{1cm} (23.20)

This proves the existence part of the theorem. To verify uniqueness, suppose $h_1$ and $h_2$ both represent $\Lambda$ and let $f \in L^1(\Omega)$ be such that $|f| \leq 1$ and $f(h_1 - h_2) = |h_1 - h_2|$. Then

$$0 = \Lambda f - \Lambda f = \int (h_1 - h_2) f d\mu = \int |h_1 - h_2| d\mu.$$  

Thus $h_1 = h_2$. Finally,

$$\|\Lambda\| = \sup\{|\int h f d\mu| : \|f\|_1 \leq 1\} \leq \|h\|_\infty \leq \|\Lambda\|$$

by 23.20. \hspace{1cm} (23.21)

Next these results are extended to the $\sigma$ finite case.

**Lemma 23.5.5** Let $(\Omega, S, \mu)$ be a measure space and suppose there exists a measurable function, $r$ such that $r(x) > 0$ for all $x$, there exists $M$ such that $|r(x)| < M$ for all $x$, and $\int r d\mu < \infty$. Then for

$$\Lambda \in (L^p(\Omega, \mu))', \ p \geq 1,$$

there exists $h \in L^q(\Omega, \mu), \ L^\infty(\Omega, \mu)$ if $p = 1$ such that

$$\Lambda f = \int h f d\mu.$$  

Also $\|h\| = \|\Lambda\|$. ($\|h\| = \|h\|_q$ if $p > 1$, $\|h\|_\infty$ if $p = 1$). Here

$$\frac{1}{p} + \frac{1}{q} = 1.$$  

**Proof:** Define a new measure $\tilde{\mu}$, according to the rule

$$\tilde{\mu}(E) = \int_E r d\mu.$$  \hspace{1cm} (23.21)

Thus $\tilde{\mu}$ is a finite measure on $S$. For $\Lambda \in (L^p(\mu))'$,

$$\Lambda(f) = \Lambda\left(r^{1/p} \left(r^{-1/p} f\right)\right) = \tilde{\Lambda}\left(r^{-1/p} f\right)$$
where
\[ \tilde{\Lambda}(g) \equiv \Lambda \left( r^{1/p} g \right) \]

Now \( \tilde{\Lambda} \) is in \( L^p(\tilde{\mu})' \) because
\[
|\tilde{\Lambda}(g)| = |\Lambda \left( r^{1/p} g \right) | \leq ||\Lambda|| \left( \int_\Omega |r^{1/p} g|^{p} d\tilde{\mu} \right)^{1/p} \]
\[
= ||\Lambda|| \left( \int_\Omega |g|^{p} r d\mu \right)^{1/p} = ||\Lambda|| ||g||_{L^p(\tilde{\mu})}
\]

Therefore, by Theorems 23.5.4 and 23.5.1 there exists a unique \( h \in L^q(\tilde{\mu}) \) which represents \( \tilde{\Lambda} \). Here \( q = \infty \) if \( p = 1 \) and satisfies \( 1/q + 1/p = 1 \) otherwise. Then
\[
\Lambda(f) = \tilde{\Lambda}(r^{-1/p} f) = \int_\Omega h f r^{-1/p} r d\mu = \int_\Omega f \left( h r^{1/q} \right) d\mu
\]

Now \( hr^{1/q} \equiv \tilde{h} \in L^q(\tilde{\mu}) \) since \( h \in L^q(\tilde{\mu}) \). In case \( p = 1, L^q(\tilde{\mu}) \) and \( L^q(\mu) \) are exactly the same. In this case you have
\[
\Lambda(f) = \tilde{\Lambda}(r^{-1} f) = \int_\Omega h f r^{-1} r d\mu = \int_\Omega f h d\mu
\]

Thus the desired representation holds. Then in any case,
\[
|\Lambda(f)| \leq ||\tilde{h}||_{L^q} ||f||_{L^p}
\]
so
\[
||\Lambda|| \leq ||\tilde{h}||_{L^q}.
\]

Also, as before,
\[
||\tilde{h}||_{L^q(\mu)}^q = \left| \int_\Omega \tilde{h} |\tilde{h}|^{q-2} \tilde{h} d\mu \right| = \Lambda \left( \left| \tilde{h} \right|^{q-2} \tilde{h} \right)
\]
\[
\leq ||\Lambda|| \left( \int_\Omega \left| \tilde{h} \right|^{q-2} \tilde{h} \right)^{1/p} \]
\[
= ||\Lambda|| \left( \int_\Omega \left| \tilde{h} \right|^{q} \right)^{1/p} = ||\Lambda|| ||h||_{L^q}^{q/p}
\]

and so
\[
||\tilde{h}||_{L^q(\mu)} \leq ||\Lambda||
\]

It works the same for \( p = 1 \). Thus \( ||\tilde{h}||_{L^q(\mu)} = ||\Lambda|| \). \( \blacksquare \)

A situation in which the conditions of the lemma are satisfied is the case where the measure space is \( \sigma \) finite. In fact, you should show this is the only case in which the conditions of the above lemma hold.

**Theorem 23.5.6 (Riesz representation theorem)** Let \( (\Omega, S, \mu) \) be \( \sigma \) finite and let
\[ \Lambda \in (L^p(\Omega, \mu))', \ p \geq 1. \]

Then there exists a unique \( h \in L^q(\Omega, \mu), \ L^\infty(\Omega, \mu) \) if \( p = 1 \) such that
\[
\Lambda f = \int h f d\mu.
\]

Also \( ||h|| = ||\Lambda||. (||h|| = ||h||_q \ if \ p > 1, \ ||h||_{\infty} \ if \ p = 1) \). Here
\[
\frac{1}{p} + \frac{1}{q} = 1.
\]
23.6. THE DUAL SPACE OF $C_0(X)$

**Proof:** Without loss of generality, assume $\mu(\Omega) = \infty$. Then let $\{\Omega_n\}$ be a sequence of disjoint elements of $S$ having the property that

$$1 < \mu(\Omega_n) < \infty, \quad \cup_{n=1}^{\infty} \Omega_n = \Omega.$$ 

Define

$$r(x) = \sum_{n=1}^{\infty} \frac{1}{n^2} \chi_{\Omega_n}(x) \mu(\Omega_n)^{-1}, \quad \tilde{\mu}(E) = \int_E r d\mu.$$ 

Thus

$$\int_{\Omega} r d\mu = \tilde{\mu}(\Omega) = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$$

so $\tilde{\mu}$ is a finite measure. The above lemma gives the existence part of the conclusion of the theorem. Uniqueness is done as before. This proves the theorem. This leads to a lot more interesting results in functional analysis.

23.6  The Dual Space Of $C_0(X)$

Consider the dual space of $C_0(X)$ where $X$ is a polish space in which the balls have compact closure. It will turn out to be a space of measures. To show this, the following lemma will be convenient. Recall $C_0(X)$ is defined as follows.

**Definition 23.6.1** $f \in C_0(X)$ means that for every $\varepsilon > 0$ there exists a compact set $K$ such that $|f(x)| < \varepsilon$ whenever $x \notin K$. Recall the norm on this space is

$$\|f\|_\infty \equiv \|f\| \equiv \sup \{|f(x)| : x \in X\}$$

The next lemma has to do with extending functionals which are defined on nonnegative functions to complex valued functions in such a way that the extended function is linear. This exact process was used earlier with the abstract Lebesgue integral. Basically, you can do it when the functional “desires to be linear”.

**Lemma 23.6.2** Suppose $\lambda$ is a mapping which has nonnegative real values which is defined on the nonnegative functions in $C_0(X)$ such that

$$\lambda(af + bg) = a\lambda(f) + b\lambda(g) \quad (23.22)$$

whenever $a, b \geq 0$ and $f, g \geq 0$. Then there exists a unique extension of $\lambda$ to all of $C_0(X)$, $\Lambda$ such that whenever $f, g \in C_0(X)$ and $a, b \in \mathbb{C}$, it follows

$$\Lambda(af + bg) = a\Lambda(f) + b\Lambda(g).$$

If

$$|\lambda(f)| \leq C \|f\|_\infty \quad (23.23)$$

then

$$|\Lambda f| \leq C \|f\|_\infty, \quad |\Lambda f| \leq \lambda(|f|).$$

**Proof:** Here $\lambda$ is defined on the nonnegative functions. First extend it to the continuous real valued functions. There is only one way to do it and retain the map is linear. Let $C_0(X; \mathbb{R})$ be the real-valued functions in $C_0(X)$ and define

$$\Lambda_R(f) = \Lambda_R(f^+ - f^-) = \Lambda_R(f^+) - \Lambda_R(f^-) = \lambda f^+ - \lambda f^-$$

for $f \in C_0(X; \mathbb{R})$. This is the only thing possible if $\Lambda_R$ is to be linear. Is $\Lambda_R(f + g) = \Lambda_R(f) + \Lambda_R(g)$?

$$f + g = (f + g)^+ - (f + g)^- = f^+ - f^- + g^+ - g^-$$
and so
\[ \Lambda_R (f + g) = \Lambda_R (f + g)^+ - \Lambda_R (f + g)^- \]
\[ \Lambda_R f + \Lambda_R g = \Lambda_R f^+ - \Lambda_R f^- + \Lambda_R g^+ - \Lambda_R g^- \]
Are these equal? This will be so if and only if
\[ \Lambda_R (f + g)^+ + \Lambda_R f^- + \Lambda_R g^- = \Lambda_R (f + g)^- + \Lambda_R f^+ + \Lambda_R g^+ \]
equivalently, since \( \Lambda_R = \lambda \) on nonnegative functions and \( \lambda \) tries to be linear,
\[ \lambda \left( (f + g)^+ + f^- + g^- \right) = \lambda \left( (f + g)^- + f^+ + g^+ \right) \]
But this is so because \((f + g)^+ + f^- + g^- = (f + g)^- + f^+ + g^+\).

It is necessary to verify that \( \Lambda_R(cf) = c\Lambda_R(f) \) for all \( c \in \mathbb{R} \). But \( (cf)^\pm = cf^\pm \), if \( c \geq 0 \) while \((cf)^- = (-c)(f)^-\), if \( c < 0 \) and \((cf)^- = (-c)(f^+)^-\), if \( c < 0 \). Thus, if \( c < 0 \),
\[ \Lambda_R(cf) = \lambda(cf)^+ - \lambda(cf)^- = \lambda((-c)(f)^-) - \lambda((-c)(f^+)^-) \]
\[ = -c\lambda(f^-) + c\lambda(f^+) = c(\lambda(f^+) - \lambda(f^-)) = c\Lambda_R(f) \].

A similar formula holds more easily if \( c \geq 0 \). Now let
\[ \Lambda f = \lambda_R(\text{Re} f) + i\lambda_R(\text{Im} f) \]
for arbitrary \( f \in C_0(X) \). This is the only possible definition if it is to be linear and agree with \( \Lambda_R \) on \( C_0(X; \mathbb{R}) \). Then
\[ \Lambda (f + g) = \Lambda_R (\text{Re}(f + g)) + i\Lambda_R (\text{Im}(f + g)) \]
\[ = \Lambda_R (\text{Re} f) + \Lambda_R (\text{Re} g) + i [\Lambda_R (\text{Im} f) + \Lambda_R (\text{Im} g)] \]
\[ = \Lambda (f) + \Lambda (g) \]
Also for \( b \) real, \( \text{Re} (ibf) = -b \text{Im} f, \text{Im} (ibf) = b \text{Re} (f) \). Then letting \( f = u + iv \)
\[ \Lambda ((a + ib)(u + iv)) = \Lambda (au - bv + iav + ibu) \]
\[ = \Lambda_R (au - bv) + i\Lambda_R (av + bu) \]
\[ (a + ib) \Lambda (u + iv) = (a + ib) (\Lambda_R (u) + i\Lambda_R (v)) \]
\[ = a\Lambda_R (u) - b\Lambda_R (v) + i [b\Lambda_R (u) + a\Lambda_R (v)] \]
which is the same thing because \( \Lambda_R \) is linear.

It remains to verify the claim about continuity of \( \Lambda \) in case of \( \mathbb{R}^2 \). This is really pretty obvious because \( f_n \to 0 \) in \( C_0(X) \) if and only if the positive and negative parts of real and imaginary parts also converge to 0 and \( \lambda \) of each of these converges to 0 by assumption. What of the last claim that \( |\Lambda f| \leq \lambda(|f|) \)? Let \( \omega \) have \( |\omega| = 1 \) and \( |\Lambda f| = \omega \Lambda (f) \). Since \( \Lambda \) is linear,
\[ |\Lambda f| = \omega \Lambda (f) = \Lambda (\omega f) = \Lambda_R (\text{Re} \omega f) \leq \Lambda_R \left( \text{Re} (\omega f)^\pm \right) = \lambda \left( \text{Re} (\omega f)^+ \right) \leq \lambda(|f|) \]

Let \( L \in C_0(X)' \). Also denote by \( C_0^+(X) \) the set of nonnegative continuous functions defined on \( X \).

**Definition 23.6.3** Letting \( L \in C_0(X)' \), define for \( f \in C_0^+(X) \)
\[ \lambda(f) = \sup \{ |Lg| : |g| \leq f \}. \]

Note that \( \lambda(f) < \infty \) because \( |Lg| \leq \|L\| |g| \leq \|L\| |f| \) for \( |g| \leq f \). Isn’t this a lot like the total variation of a vector measure? Indeed it is, and the proof that \( \lambda \) wants to be linear is also similar to the proof that the total variation is a measure. This is the content of the following lemma.
Lemma 23.6.4 If \( c \geq 0 \), \( \lambda(cf) = c\lambda(f) \), \( f_1 \leq f_2 \) implies \( \lambda f_1 \leq \lambda f_2 \), and 
\[
\lambda(f_1 + f_2) = \lambda(f_1) + \lambda(f_2).
\]
Also
\[
0 \leq \lambda(f) \leq \|L\| \|f\|_\infty
\]

**Proof:** The first two assertions are easy to see so consider the third. For \( f_j \in C_0^+ (X) \), there exists \( g_i \in C_0 (X) \) such that \( |g_i| \leq f_i \) and
\[
\lambda(f_1) + \lambda(f_2) \leq |L(g_1)| + |L(g_2)| + 2\varepsilon
\]
\[
= L(\omega_1 g_1) + L(\omega_2 g_2) + 2\varepsilon
\]
\[
= L(\omega_1 g_1 + \omega_2 g_2) + 2\varepsilon
\]
\[
= |L(\omega_1 g_1 + \omega_2 g_2)| + 2\varepsilon
\]
where \( |g_i| \leq f_i \) and \( |\omega_i| = 1 \) and \( \omega_i L(g_i) = |L(g_i)| \). Now
\[
|\omega_1 g_1 + \omega_2 g_2| \leq |g_1| + |g_2| \leq f_1 + f_2
\]
and so the above shows
\[
\lambda(f_1) + \lambda(f_2) \leq \lambda(f_1 + f_2) + 2\varepsilon.
\]
Since \( \varepsilon \) is arbitrary, \( \lambda(f_1) + \lambda(f_2) \leq \lambda(f_1 + f_2) \). It remains to verify the other inequality.

Now let \( |g| \leq f_1 + f_2 \),
\[
|Lg| \geq \lambda(f_1 + f_2) - \varepsilon.
\]
Let
\[
h_i (x) = \begin{cases} 
\frac{f_i(x)g(x)}{f_1(x) + f_2(x)} & \text{if } f_1 (x) + f_2 (x) > 0, \\
0 & \text{if } f_1 (x) + f_2 (x) = 0.
\end{cases}
\]
Then \( h_i \) is continuous and
\[
h_1(x) + h_2(x) = g(x), \ |h_i| \leq f_i.
\]
The function \( h_i \) is clearly continuous at points \( x \) where \( f_1 (x) + f_2 (x) > 0 \). The reason it is continuous at a point where \( f_1 (x) + f_2 (x) = 0 \) is that at every point \( y \) where \( f_1 (y) + f_2 (y) > 0 \), the top description of the function gives
\[
|h_i (y)| = \left| \frac{f_i(y) g(y)}{f_1 (y) + f_2 (y)} \right| \leq |g(y)| \leq f_1 (y) + f_2 (y)
\]
so if \( |y - x| \) is small enough, \( |h_i (y)| \) is also small. Then it follows from this definition of the \( h_i \) that
\[
-\varepsilon + \lambda(f_1 + f_2) \leq |Lg| = |Lh_1 + Lh_2| \leq |Lh_1| + |Lh_2|
\]
\[
\leq \lambda(f_1) + \lambda(f_2).
\]
Since \( \varepsilon > 0 \) is arbitrary, this shows that
\[
\lambda(f_1 + f_2) \leq \lambda(f_1) + \lambda(f_2) \leq \lambda(f_1 + f_2)
\]
The last assertion follows from
\[
\lambda(f) = \sup\{|Lg| : |g| \leq f\} \leq \sup_{\|g\|_\infty \leq \|f\|_\infty} \|L\| \|g\|_\infty \leq \|L\| \|f\|_\infty.
\]
Let \( \Lambda \) be defined in Lemma 23.5. We Then \( \Lambda \) is linear by this lemma and also satisfies
\[
|\Lambda f| = \omega \Lambda f = \Lambda(\omega f) \leq \lambda(|f|) \leq \|L\| \|f\|_\infty.
\]
Also, if \( f \geq 0 \),
\[
\Lambda f = \Lambda R f = \lambda(f) \geq 0.
\]
Therefore, \( \Lambda \) is a positive linear functional on \( C_0(X) \). In particular, it is a positive linear functional on \( C_c(X) \). Thus there are now two linear continuous mappings \( L, \Lambda \) which are defined on \( C_0(X) \).

The above shows that in fact \( \| \Lambda \| \leq \| L \| \). Also, from the definition of \( \Lambda \)
\[
|Lg| \leq \lambda (|g|) = \Lambda (|g|) \leq \| \Lambda \| \| g \|_{\infty}
\]
so in fact, \( \| L \| \leq \| \Lambda \| \) showing that these two have the same operator norms.

\[
\| L \| = \| \Lambda \| \tag{23.25}
\]

By Theorem 21.0.5 on Page 506, there exists a unique measure \( \mu \) such that
\[
\lambda f = \int_X f \, d\mu
\]
for all \( f \in C_c(X) \). This measure is regular. In fact, it is actually a finite measure.

**Lemma 23.6.5** Let \( L \in C_0(X)' \) as above. Then letting \( \mu \) be the Radon measure just described, it follows \( \mu \) is finite and
\[
\mu(X) = \| \Lambda \| = \| L \|
\]

**Proof:** First of all, it was observed above in 23.25 that \( \| \Lambda \| = \| L \| \).

Now by regularity,
\[
\mu(X) = \sup \{ \mu(K) : K \subset X \}
\]
and so letting \( K \prec f \prec X \) for one of these \( K \), it also follows
\[
\mu(X) = \sup \{ \lambda f : f \prec X \}
\]

However, for such nonnegative \( f \prec X \),
\[
\lambda(f) \equiv \sup \{ |Lg| : |g| \leq f \} \leq \sup \{ \| L \| \| g \|_{\infty} : |g| \leq f \}
\]
and so
\[
0 \leq \lambda f = \lambda f \leq \| L \|
\]
It follows that
\[
\mu(X) = \sup \{ \lambda f : f \prec X \} \leq \| L \|.
\]
Now since \( C_c(X) \) is dense in \( C_0(X) \), there exists \( f \in C_c(X) \) such that \( \| f \|_{\infty} \leq 1 \) and
\[
|\lambda f + \varepsilon > \lambda f| \leq \| L \| = \| \Lambda \| = \| L \|
\]
Thus,
\[
\| L \| - \varepsilon < |\lambda f| \leq \lambda (|f|) = \Lambda |f| \leq \mu(X)
\]
Since \( \varepsilon \) is arbitrary, this shows \( \| L \| = \mu(X) \). □

What follows is the Riesz representation theorem for \( C_0(X)' \).

**Theorem 23.6.6** Let \( L \in (C_0(X))' \) for \( X \) a Polish space with closed balls compact. Then there exists a finite Radon measure \( \mu \) and a function \( \sigma \in L^\infty(X, \mu) \) such that for all \( f \in C_0(X) \),
\[
L(f) = \int_X f \sigma d\mu.
\]
Furthermore,
\[
\mu(X) = \| L \|, \ |\sigma| = 1 \ a.e.
\]
and if
\[
\nu(E) \equiv \int_E \sigma d\mu
\]
then \( \mu = |\nu| \).
Proof: From the above there exists a unique Radon measure \( \mu \) such that for all \( f \in C_c(X) \),

\[
\Lambda f = \int_X f d\mu
\]

Then for \( f \in C_c(X) \),

\[
|Lf| \leq \lambda (|f|) = \Lambda(|f|) = \int_X |f| d\mu = \|f\|_{L^1(\mu)}.
\]

Since \( \mu \) is both inner and outer regular, \( C_c(X) \) is dense in \( L^1(X, \mu) \). (See Theorem \ref{thm:density} for more than is needed.) Therefore \( L \) extends uniquely to an element of \( (L^1(X, \mu))^\prime, L. \) By the Riesz representation theorem for \( L^1 \) for finite measure spaces, there exists a unique \( \sigma \in L^\infty(X, \mu) \) such that for all \( f \in L^1(X, \mu) \),

\[
\tilde{L} f = \int_X f \sigma d\mu
\]

In particular, for all \( f \in C_0(X) \),

\[
Lf = \int_X f \sigma d\mu
\]

and it follows from Lemma \ref{lem:existence}, \( \mu(X) = \|L\| \).

It remains to verify \( |\sigma| = 1 \) a.e. For any continuous \( f \geq 0 \),

\[
\Lambda f \equiv \int_X f d\mu \geq |Lf| = \left| \int_X f \sigma d\mu \right|
\]

Now if \( E \) is measurable, the regularity of \( \mu \) implies there exists a sequence of nonnegative bounded functions \( f_n \in C_c(X) \) such that \( f_n(x) \to X_E (x) \) a.e. and in \( L^1(\mu) \). Then using the dominated convergence theorem in the above,

\[
\int_E d\mu = \lim_{n \to \infty} \int_X f_n d\mu = \lim_{n \to \infty} \Lambda(f_n) \geq \lim_{n \to \infty} |Lf_n| = \lim_{n \to \infty} \left| \int_X f_n \sigma d\mu \right|
\]

and so if \( \mu(E) > 0 \),

\[
1 \geq \left| \frac{1}{\mu(E)} \int_E \sigma d\mu \right|
\]

which shows from Lemma \ref{lem:existence} that \( |\sigma| \leq 1 \) a.e. But also, choosing \( f_1 \) appropriately, \( \|f_1\|_\infty \leq 1 \),

\[
|Lf_1| + \varepsilon > \|L\| = \mu(X)
\]

Letting \( \omega (Lf_1) = |Lf_1|, |\omega| = 1 \),

\[
\mu(X) = \|L\| = \sup_{\|f\|_\infty \leq 1} |Lf| \leq |Lf_1| + \varepsilon
\]

\[
= \omega Lf_1 + \varepsilon = \int_X f_1 \omega \sigma d\mu + \varepsilon
\]

\[
= \int_X \text{Re} (f_1 \omega \sigma) d\mu + \varepsilon
\]

\[
\leq \int_X |\sigma| d\mu + \varepsilon
\]

and since \( \varepsilon \) is arbitrary,

\[
\mu(X) \leq \int_X |\sigma| d\mu \leq \mu(X)
\]

which requires \( |\sigma| = 1 \) a.e. since it was shown to be no larger than 1 and if it is smaller than 1 on a set of positive measure, then the above could not hold.
It only remains to verify \( \mu = |\nu| \). By Corollary 23.4.9,
\[
|\nu|(E) = \int_E |\sigma| \, d\mu = \int_E 1 \, d\mu = \mu(E)
\]
and so \( \mu = |\nu| \). \( \square \)

Sometimes people write
\[
\int_X fd\nu \equiv \int_X f|\sigma| \, d|\nu|
\]
where \( \sigma |\nu| \) is the polar decomposition of the complex measure \( \nu \). Then with this convention, the above representation is
\[
L(f) = \int_X f d\nu, \quad |\nu|(X) = ||L||.
\]
Also note that at most one \( \nu \) can represent \( L \). If there were two of them \( \nu_1, i = 1, 2 \), then \( \nu_1 - \nu_2 \) would represent 0 and so \( |\nu_1 - \nu_2|(X) = 0 \). Hence \( \nu_1 = \nu_2 \).

### 23.7 Exercises

1. Suppose \( \mu \) is a vector measure having values in \( \mathbb{R}^n \) or \( \mathbb{C}^n \). Can you show that \( |\mu| \) must be finite? **Hint:** You might define for each \( e_i \), one of the standard basis vectors, the real or complex measure, \( \mu_{e_i} \) given by \( \mu_{e_i}(E) = e_i \cdot \mu(E) \). Why would this approach not yield anything for an infinite dimensional normed linear space in place of \( \mathbb{R}^n \)?

2. The Riesz representation theorem of the \( L^p \) spaces can be used to prove a very interesting inequality. Let \( r, p, q \in (1, \infty) \) satisfy
\[
\frac{1}{r} = \frac{1}{p} + \frac{1}{q} - 1.
\]
Then
\[
\frac{1}{q} = 1 + \frac{1}{r} - \frac{1}{p} > \frac{1}{r}
\]
and so \( r > q \). Let \( \theta \in (0, 1) \) be chosen so that \( \theta r = q \). Then also
\[
\frac{1}{r} = \left( \frac{1/p+1/p'=1}{1 - \frac{1}{p'}} \right) + \frac{1}{q} - 1 = \frac{1}{q} - \frac{1}{p'}
\]
and so
\[
\frac{\theta}{q} = \frac{1}{q} - \frac{1}{p'}
\]
which implies \( p' (1 - \theta) = q \). Now let \( f \in L^p(\mathbb{R}^n) \), \( g \in L^q(\mathbb{R}^n) \), \( f, g \geq 0 \). Justify the steps in the following argument using what was just shown that \( \theta r = q \) and \( p' (1 - \theta) = q \). Let
\[
h \in L^{p'}(\mathbb{R}^n). \quad \left( \frac{1}{r} + \frac{1}{p'} = 1 \right)
\]
\[
\left| \int f * g \, h \, dx \right| = \left| \int \int f(\mathbf{y}) g(\mathbf{x} - \mathbf{y}) h(\mathbf{x}) \, dy \, dx \right|.
\]
\[
\leq \int \int |f(\mathbf{y})| |g(\mathbf{x} - \mathbf{y})|^{\theta} |h(\mathbf{x})|^{1-\theta} \, dy \, dx
\]
\[
\leq \int \left( \int \left( |g(\mathbf{x} - \mathbf{y})|^{1-\theta} |h(\mathbf{x})| \right)^{p'} \, dy \right)^{1/p'} \, dx.
\]
Suppose \( \Omega \) continues to hold if \( r, p, q \) are only assumed to lie in \([1, \infty]\). The \( \|f \ast g\|_r \leq \|g\|_q \|f\|_p \). How does this inequality follow from the above computation? Does it continue to hold if \( r, p, q \) are only assumed to be in \([1, \infty]\)? Explain. Does it hold even if \( r, p, q \) are only assumed to lie in \([1, \infty]\)?

3. Suppose \( (\Omega, \mu, S) \) is a finite measure space and that \( \{f_n\} \) is a sequence of functions which converge weakly to 0 in \( L^p(\Omega) \). This means that

\[
\int_{\Omega} f_n g d\mu \to 0
\]

for every \( g \in L^{p'}(\Omega) \). Suppose also that \( f_n(x) \to 0 \) a.e. Show that then \( f_n \to 0 \) in \( L^{p-\varepsilon}(\Omega) \) for every \( p > \varepsilon > 0 \).

4. Give an example of a sequence of functions in \( L^\infty(-\pi, \pi) \) which converges weak * to zero but which does not converge pointwise a.e. to zero. Convergence weak * to 0 means that for every \( g \in L^1(-\pi, \pi), \int_{-\pi}^{\pi} g(t) f_n(t) dt \to 0 \). \textbf{Hint:} First consider \( g \in C_c^\infty(-\pi, \pi) \) and maybe try something like \( f_n(t) = \sin(nt) \). Do integration by parts.

5. Let \( \lambda \) be a real vector measure on the measure space \( (\Omega, F) \). That is \( \lambda \) has values in \( \mathbb{R} \). The Hahn decomposition says there exist measurable sets \( P, N \) such that

\[
\lambda(P \cup N) = \Omega, \lambda(P \cap N) = 0,
\]

and for each \( F \subseteq P, \lambda(F) \geq 0 \) and for each \( F \subseteq N, \lambda(F) \leq 0 \). These sets \( P, N \) are called the positive set and the negative set respectively. Show the existence of the Hahn decomposition. Also explain how this decomposition is unique in the sense that if \( P', N' \) is another Hahn decomposition, then \( (P \setminus P') \cup (P' \setminus P) \) has measure zero, a similar formula holding for \( N, N' \). When you have the Hahn decomposition, as just described, you define \( \lambda^+(E) \equiv \lambda(E \cap P), \lambda^-(E) \equiv \lambda(E \cap N) \). This is sometimes called the Hahn Jordan decomposition. \textbf{Hint:} This is pretty easy if you use the polar decomposition above.
6. The Hahn decomposition holds for measures which have values in \((-\infty, \infty]\). Let \(\lambda\) be such a measure which is defined on a \(\sigma\) algebra of sets \(\mathcal{F}\). This is not a vector measure because the set on which it has values is not a vector space. Thus this case is not included in the above discussion. \(N \in \mathcal{F}\) is called a negative set if \(\lambda(B) \leq 0\) for all \(B \subseteq N\). \(P \in \mathcal{F}\) is called a positive set if for all \(F \subseteq P, \lambda(F) \geq 0\). (Here it is always assumed you are only considering sets of \(\mathcal{F}\).) Show that if \(\lambda(A) \leq 0\), then there exists \(N \subseteq A\) such that \(N\) is a negative set and \(\lambda(N) \leq \lambda(A)\). **Hint:** This is done by subtracting off disjoint sets having positive measure. Let \(A \equiv N_0\) and suppose \(N_n \subseteq A\) has been obtained. Tell why

\[
t_n = \sup \{\lambda(E) : E \subseteq N_n\} \geq 0.
\]

Let \(B_n \subseteq N_n\) such that

\[
\lambda(B_n) > \frac{t_n}{2}
\]

Then \(N_{n+1} = N_n \setminus B_n\). Thus the \(N_n\) are decreasing in \(n\) and the \(B_n\) are disjoint. Explain why \(\lambda(N_n) \leq \lambda(N_0)\). Let \(N = \cap N_n\). Argue \(t_n\) must converge to 0 since otherwise \(\lambda(N) = -\infty\). Explain why this requires \(N\) to be a negative set in \(A\) which has measure no larger than that of \(A\).

7. Using Problem 9 complete the Hahn decomposition for \(\lambda\) having values in \((-\infty, \infty]\). Now the Hahn Jordan decomposition for the measure \(\lambda\) is

\[
\lambda^+ (E) \equiv \lambda(E \cap P), \quad \lambda^- (E) \equiv -\lambda(E \cap N).
\]

Explain why \(\lambda^-\) is a finite measure. **Hint:** Let \(N_0 = 0\). For \(N_n\) a given negative set, let

\[
t_n = \inf \{\lambda(E) : E \cap N_n = \emptyset\}
\]

Explain why you can assume that for all \(n, t_n < 0\). Let \(E_n \subseteq N_n^c\) such that

\[
\lambda(E_n) < t_n/2 < 0
\]

and from Problem 9 let \(A_n \subseteq E_n\) be a negative set such that \(\lambda(A_n) \leq \lambda(E_n)\). Then \(N_{n+1} = N_n \cup A_n\). If \(t_n\) does not converge to 0 explain why there exists a set having measure \(-\infty\) which is not allowed. Thus \(t_n \to 0\). Let \(N = \cup_{n=1}^{\infty} N_n\) and explain why \(P = N^c\) must be positive due to \(t_n \to 0\).

8. What if \(\lambda\) has values in \([-\infty, \infty]\). Prove there exists a Hahn decomposition for \(\lambda\) as in the above problem. Why do we not allow \(\lambda\) to have values in \([-\infty, \infty]\)? **Hint:** You might want to consider \(-\lambda\).

9. Suppose \(X\) is a Banach space and let \(X'\) denote its dual space. A sequence \((x_n^*)\) in \(X'\) is said to converge weak * to \(x^* \in X'\) if for every \(x \in X\),

\[
\lim_{n \to \infty} x_n^*(x) = x^*(x).
\]

Let \(\phi_n\) be a mollifier. Also let \(\delta\) be the measure defined by

\[
\delta(E) = 1 \text{ if } 0 \in E \text{ and } 0 \text{ if } 1 \notin E.
\]

Explain how \(\phi_n \to \delta\) weak *.

10. A Banach space \(X\) is called strictly convex if \(\|\frac{x+y}{2}\| < \frac{1}{2} \|x\| + \frac{1}{2} \|y\|\). A Banach space is called uniformly convex if whenever \(\|x_n + y_n\| \to 2\), it follows that \(\|x_n - y_n\| \to 0\). Show that uniform convexity implies strict convexity. It was not done here, but it can be proved using something called Clarkson’s inequalities that the \(L^p\) spaces for \(p > 1\) are uniformly convex.
Part IV

Appendix
Appendix A

The Cross Product

This short explanation is included for the sake of those who have had a calculus course in which the geometry of the cross product has not been made clear. Unfortunately, this is the case in most of the current books. The geometric significance in terms of angles and lengths is very important, not just a formula for computing the cross product.

The cross product is the other way of multiplying two vectors in $\mathbb{R}^3$. It is very different from the dot product in many ways. First the geometric meaning is discussed and then a description in terms of coordinates is given. Both descriptions of the cross product are important. The geometric description is essential in order to understand the applications to physics and geometry while the coordinate description is the only way to practically compute the cross product.

**Definition A.0.1** Three vectors, $\mathbf{a}, \mathbf{b}, \mathbf{c}$ form a right handed system if when you extend the fingers of your right hand along the vector $\mathbf{a}$ and close them in the direction of $\mathbf{b}$, the thumb points roughly in the direction of $\mathbf{c}$.

For an example of a right handed system of vectors, see the following picture.

In this picture the vector $\mathbf{c}$ points upwards from the plane determined by the other two vectors. You should consider how a right hand system would differ from a left hand system. Try using your left hand and you will see that the vector $\mathbf{c}$ would need to point in the opposite direction as it would for a right hand system.

From now on, the vectors, $\mathbf{i}, \mathbf{j}, \mathbf{k}$ will always form a right handed system. To repeat, if you extend the fingers of your right hand along $\mathbf{i}$ and close them in the direction $\mathbf{j}$, the thumb points in the direction of $\mathbf{k}$. The following is the geometric description of the cross product. It gives both the direction and the magnitude and therefore specifies the vector.

**Definition A.0.2** Let $\mathbf{a}$ and $\mathbf{b}$ be two vectors in $\mathbb{R}^3$. Then $\mathbf{a} \times \mathbf{b}$ is defined by the following two rules.

1. $|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}| |\mathbf{b}| \sin \theta$ where $\theta$ is the included angle.
2. \( \mathbf{a} \times \mathbf{b} \cdot \mathbf{a} = 0, \mathbf{a} \times \mathbf{b} \cdot \mathbf{b} = 0, \) and \( \mathbf{a}, \mathbf{b}, \mathbf{a} \times \mathbf{b} \) forms a right hand system.

Note that \( |\mathbf{a} \times \mathbf{b}| \) is the area of the parallelogram determined by \( \mathbf{a} \) and \( \mathbf{b} \).

The cross product satisfies the following properties.

\[
\mathbf{a} \times \mathbf{b} = -(\mathbf{b} \times \mathbf{a}), \quad \mathbf{a} \times \mathbf{a} = 0 \tag{1.1}
\]

For \( \alpha \) a scalar,

\[
(\alpha \mathbf{a}) \times \mathbf{b} = \alpha (\mathbf{a} \times \mathbf{b}) = \mathbf{a} \times (\alpha \mathbf{b}) \tag{1.2}
\]

For \( \mathbf{a}, \mathbf{b}, \) and \( \mathbf{c} \) vectors, one obtains the distributive laws,

\[
\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}, \tag{1.3}
\]

\[
(\mathbf{b} + \mathbf{c}) \times \mathbf{a} = \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a} \tag{1.4}
\]

Formula (1.1) follows immediately from the definition. The vectors \( \mathbf{a} \times \mathbf{b} \) and \( \mathbf{b} \times \mathbf{a} \) have the same magnitude, \( |\mathbf{a}| |\mathbf{b}| \sin \theta \), and an application of the right hand rule shows they have opposite direction. Formula (1.2) is also fairly clear. If \( \alpha \) is a nonnegative scalar, the direction of \( (\alpha \mathbf{a}) \times \mathbf{b} \) is the same as the direction of \( \mathbf{a} \times \mathbf{b} \), \( \alpha (\mathbf{a} \times \mathbf{b}) \) and \( \mathbf{a} \times (\alpha \mathbf{b}) \) while the magnitude is just \( \alpha \) times the magnitude of \( \mathbf{a} \times \mathbf{b} \) which is the same as the magnitude of \( \mathbf{a} \times \mathbf{b} \). Using this yields equality in (1.2). In the case where \( \alpha < 0 \), everything works the same way except the vectors are all pointing in the opposite direction and you must multiply by \(|\alpha| \) when comparing their magnitudes. The distributive laws are much harder to establish but the second follows from the first quite easily. Thus, assuming the first, and using (1.1)

\[
(\mathbf{b} + \mathbf{c}) \times \mathbf{a} = -\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = -(\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c})
\]

\[
= \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}.
\]

A proof of the distributive law is given in a later section for those who are interested.

Now from the definition of the cross product,

\[
i \times j = k \quad j \times i = -k
\]

\[
k \times i = j \quad i \times k = -j
\]

\[
j \times k = i \quad k \times j = -i
\]

With this information, the following gives the coordinate description of the cross product.

**Proposition A.0.3** Let \( \mathbf{a} = a_1 \mathbf{i} + a_2 \mathbf{j} + a_3 \mathbf{k} \) and \( \mathbf{b} = b_1 \mathbf{i} + b_2 \mathbf{j} + b_3 \mathbf{k} \) be two vectors. Then

\[
\mathbf{a} \times \mathbf{b} = (a_2 b_3 - a_3 b_2) \mathbf{i} + (a_3 b_1 - a_1 b_3) \mathbf{j} + (a_1 b_2 - a_2 b_1) \mathbf{k} \tag{1.5}
\]

**Proof:** From the above table and the properties of the cross product listed,

\[
(a_1 \mathbf{i} + a_2 \mathbf{j} + a_3 \mathbf{k}) \times (b_1 \mathbf{i} + b_2 \mathbf{j} + b_3 \mathbf{k}) =
\]

\[
a_1 b_2 \mathbf{i} \times \mathbf{j} + a_1 b_3 \mathbf{i} \times \mathbf{k} + a_2 b_1 \mathbf{j} \times \mathbf{i} + a_2 b_3 \mathbf{j} \times \mathbf{k} + a_3 b_1 \mathbf{k} \times \mathbf{i} + a_3 b_2 \mathbf{k} \times \mathbf{j}
\]
\[ +a_3b_1k \times i + a_3b_2k \times j \]
\[ = a_1b_2k - a_1b_3j - a_2b_1k + a_2b_3i + a_3b_1j - a_3b_2i \]
\[ = (a_2b_3 - a_3b_2)i + (a_3b_1 - a_1b_3)j + (a_1b_2 - a_2b_1)k \] (1.6)

It is probably impossible for most people to remember (1.6). Fortunately, there is a somewhat easier way to remember it. Define the determinant of a 2 \times 2 matrix as follows

\[
\begin{vmatrix}
  a & b \\
  c & d \\
\end{vmatrix} = ad - bc
\]

Then

\[
\mathbf{a} \times \mathbf{b} = \begin{vmatrix}
  i & j & k \\
  a_1 & a_2 & a_3 \\
  b_1 & b_2 & b_3 \\
\end{vmatrix}
\] (1.7)

where you expand the determinant along the top row. This yields

\[
i (-1)^{1+1} \begin{vmatrix}
  a_2 & a_3 \\
  b_2 & b_3 \\
\end{vmatrix} + j (-1)^{2+1} \begin{vmatrix}
  a_1 & a_3 \\
  b_1 & b_3 \\
\end{vmatrix} + k (-1)^{3+1} \begin{vmatrix}
  a_1 & a_2 \\
  b_1 & b_2 \\
\end{vmatrix}
\]

\[
= i \begin{vmatrix}
  a_2 & a_3 \\
  b_2 & b_3 \\
\end{vmatrix} - j \begin{vmatrix}
  a_1 & a_3 \\
  b_1 & b_3 \\
\end{vmatrix} + k \begin{vmatrix}
  a_1 & a_2 \\
  b_1 & b_2 \\
\end{vmatrix}
\]

Note that to get the scalar which multiplies \(i\) you take the determinant of what is left after deleting the first row and the first column and multiply by \((-1)^{1+1}\) because \(i\) is in the first row and the first column. Then you do the same thing for the \(j\) and \(k\). In the case of the \(j\) there is a minus sign because \(j\) is in the first row and the second column and so\((-1)^{1+2} = -1\) while the \(k\) is multiplied by \((-1)^{3+1} = 1\). The above equals

\[
(a_2b_3 - a_3b_2)i - (a_1b_3 - a_3b_1)j + (a_1b_2 - a_2b_1)k \] (1.8)

which is the same as (1.6). There will be much more presented on determinants later. For now, consider this an introduction if you have not seen this topic.

**Example A.0.4** Find \((i - j + 2k) \times (3i - 2j + k)\).

Use (1.7) to compute this.

\[
\begin{vmatrix}
  i & j & k \\
  1 & -1 & 2 \\
  3 & -2 & 1 \\
\end{vmatrix} = \begin{vmatrix}
  -1 & 2 \\
  -2 & 1 \\
\end{vmatrix}i - \begin{vmatrix}
  2 & 1 \\
  3 & 1 \\
\end{vmatrix}j + \begin{vmatrix}
  1 & -1 \\
  3 & -2 \\
\end{vmatrix}k = 3i + 5j + k.
\]

**Example A.0.5** Find the area of the parallelogram determined by the vectors,

\((i - j + 2k), (3i - 2j + k)\).

These are the same two vectors in Example A.0.4.

From Example A.0.4 and the geometric description of the cross product, the area is just the norm of the vector obtained in Example A.0.4. Thus the area is \(\sqrt{9 + 25 + 1} = \sqrt{35}\).

**Example A.0.6** Find the area of the triangle determined by \((1, 2, 3), (0, 2, 5), (5, 1, 2)\).
This triangle is obtained by connecting the three points with lines. Picking \((1, 2, 3)\) as a starting point, there are two displacement vectors, \((-1, 0, 2)\) and \((4, -1, -1)\) such that the given vector added to these displacement vectors gives the other two vectors. The area of the triangle is half the area of the parallelogram determined by \((-1, 0, 2)\) and \((4, -1, -1)\). Thus \((-1, 0, 2) \times (4, -1, -1) = (2, 7, 1)\) and so the area of the triangle is \(\frac{1}{2} \sqrt{4 + 49 + 1} = \frac{3}{2} \sqrt{6}\).

**Observation A.0.7** In general, if you have three points (vectors) in \(\mathbb{R}^3\), \(P, Q, R\) the area of the triangle is given by

\[
\frac{1}{2} |(Q - P) \times (R - P)|.
\]

### A.1 The Box Product

**Definition A.1.1** A parallelepiped determined by the three vectors, \(a, b,\) and \(c\) consists of

\[
\{ ra + sb + tc : r, s, t \in [0, 1] \}.
\]

That is, if you pick three numbers, \(r, s, t\) each in \([0, 1]\) and form \(ra + sb + tc\), then the collection of all such points is what is meant by the parallelepiped determined by these three vectors.

The following is a picture of such a thing. You notice the area of the base of the parallelepiped, the parallelogram determined by the vectors, \(a\) and \(b\) has area equal to \(|a \times b|\) while the altitude of the parallelepiped is \(|c| \cos \theta\) where \(\theta\) is the angle shown in the picture between \(c\) and \(a \times b\). Therefore, the volume of this parallelepiped is the area of the base times the altitude which is just

\[|a \times b||c| \cos \theta = a \times b \cdot c.\]

This expression is known as the box product and is sometimes written as \([a, b, c]\). You should consider what happens if you interchange the \(b\) with the \(c\) or the \(a\) with the \(c\). You can see geometrically from drawing pictures that this merely introduces a minus sign. In any case the box product of three vectors always equals either the volume of the parallelepiped determined by the three vectors or else minus this volume.

**Example A.1.2** Find the volume of the parallelepiped determined by the vectors, \(i + 2j - 5k, i + 3j - 6k, 3i + 2j + 3k\).

According to the above discussion, pick any two of these, take the cross product and then take the dot product of this with the third of these vectors. The result will be either the desired volume or minus the desired volume.

\[(i + 2j - 5k) \times (i + 3j - 6k) = \begin{vmatrix} i & j & k \\ 1 & 2 & -5 \\ 1 & 3 & -6 \end{vmatrix} = 3i + j + k\]

Now take the dot product of this vector with the third which yields

\[(3i + j + k) \cdot (3i + 2j + 3k) = 9 + 2 + 3 = 14.\]
This shows the volume of this parallelepiped is 14 cubic units.

There is a fundamental observation which comes directly from the geometric definitions of the cross product and the dot product.

**Lemma A.1.3** Let \( \mathbf{a}, \mathbf{b}, \) and \( \mathbf{c} \) be vectors. Then \( (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) \).

**Proof:** This follows from observing that either \( (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} \) and \( \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) \) both give the volume of the parallelepiped or they both give \(-1\) times the volume. ■

**Notation A.1.4** The box product \( \mathbf{a} \times \mathbf{b} \cdot \mathbf{c} = \mathbf{a} \cdot \mathbf{b} \times \mathbf{c} \) is denoted more compactly as \([\mathbf{a}, \mathbf{b}, \mathbf{c}]\).

### A.2 The Distributive Law For Cross Product

Here is a proof of the distributive law for the cross product. Let \( \mathbf{x} \) be a vector. From the above observation,

\[
\mathbf{x} \cdot \mathbf{a} \times (\mathbf{b} + \mathbf{c}) = (\mathbf{x} \times \mathbf{a}) \cdot (\mathbf{b} + \mathbf{c}) = (\mathbf{x} \times \mathbf{a}) \cdot \mathbf{b} + (\mathbf{x} \times \mathbf{a}) \cdot \mathbf{c}
\]

\[
= \mathbf{x} \cdot \mathbf{a} \times \mathbf{b} + \mathbf{x} \cdot \mathbf{a} \times \mathbf{c} = \mathbf{x} \cdot (\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}).
\]

Therefore,

\[
\mathbf{x} \cdot [\mathbf{a} \times (\mathbf{b} + \mathbf{c}) - (\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c})] = 0
\]

for all \( \mathbf{x} \). In particular, this holds for \( \mathbf{x} = \mathbf{a} \times (\mathbf{b} + \mathbf{c}) - (\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}) \) showing that \( \mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c} \) and this proves the distributive law for the cross product.
Appendix B

Weierstrass Approximation Theorem

An arbitrary continuous function defined on an interval can be approximated uniformly by a polynomial, there exists a similar theorem which is just a generalization of this which will hold for continuous functions defined on a box or more generally a closed and bounded set. However, we will settle for the case of a box first. The proof is based on the following lemma.

Lemma B.0.1 The following estimate holds for \( x \in [0, 1] \) and \( m \geq 2 \).

\[
\sum_{k=0}^{m} \binom{m}{k} (k - mx)^2 x^k (1 - x)^{m-k} \leq \frac{1}{4} m
\]

Proof: First of all, from the binomial theorem

\[
\sum_{k=0}^{m} \binom{m}{k} (tx)^k (1 - x)^{m-k} = (1 - x + tx)^m
\]

Take a derivative and then let \( t = 1 \).

\[
\sum_{k=0}^{m} \binom{m}{k} k (tx)^{k-1} x (1 - x)^{m-k} = mx (tx - x + 1)^{m-1}
\]

Then also,

\[
\sum_{k=0}^{m} \binom{m}{k} k (x)^{k} (1 - x)^{m-k} = mx
\]

Take another time derivative of both sides.

\[
\sum_{k=0}^{m} \binom{m}{k} k^2 (tx)^{k-1} x (1 - x)^{m-k} = mx \left( (tx - x + 1)^{m-1} - tx (tx - x + 1)^{m-2} + mtx (tx - x + 1)^{m-2} \right)
\]

Plug in \( t = 1 \).

\[
\sum_{k=0}^{m} \binom{m}{k} k^2 x^k (1 - x)^{m-k} = mx (mx - x + 1)
\]
Then it follows
\[ \sum_{k=0}^{m} \binom{m}{k} (k - mx)^2 x^k (1 - x)^{m-k} \]
\[ = \sum_{k=0}^{m} \binom{m}{k} (k^2 - 2kmx + x^2 m^2) x^k (1 - x)^{m-k} \]
and from what was just shown, this equals
\[ x^2 m^2 - x^2 m - mx + mx + x^2 m^2 = -x^2 m + mx = \frac{m}{4} - m \left( x - \frac{1}{2} \right)^2. \]
Thus the expression is maximized when \( x = 1/2 \) and yields \( m/4 \) in this case. This proves the lemma.

With this preparation, here is the first version of the Weierstrass approximation theorem.

**Theorem B.0.2** Let \( f \in C ([0, 1]) \) and let
\[ p_m(x) = \sum_{k=0}^{m} \binom{m}{k} x^k (1-x)^{m-k} f \left( \frac{k}{m} \right). \]
Then these polynomials converge uniformly to \( f \) on \([0, 1] \).

**Proof:** Let \( ||f||_{\infty} \) denote the largest value of \( |f| \). By uniform continuity of \( f \), there exists a \( \delta > 0 \) such that if \( |x - x'| < \delta \), then \( |f(x) - f(x')| < \varepsilon/2 \). By the binomial theorem,
\[ |p_m(x) - f(x)| \leq \sum_{k=0}^{m} \binom{m}{k} x^k (1-x)^{m-k} |f \left( \frac{k}{m} \right) - f(x)| \]
\[ \leq \sum_{\frac{k}{m} - x < \delta} \binom{m}{k} x^k (1-x)^{m-k} \left| f \left( \frac{k}{m} \right) - f(x) \right| + \]
\[ 2 \|f\|_{\infty} \sum_{\frac{k}{m} - x \geq \delta} \binom{m}{k} x^k (1-x)^{m-k} \]
Therefore,
\[ \leq \sum_{k=0}^{m} \binom{m}{k} x^k (1-x)^{m-k} \frac{\varepsilon}{2} + 2 \|f\|_{\infty} \sum_{(k-mx)^2 \geq m^2 \delta^2} \binom{m}{k} x^k (1-x)^{m-k} \]
\[ \leq \frac{\varepsilon}{2} + 2 \|f\|_{\infty} \frac{1}{m^2 \delta^2} \sum_{k=0}^{m} \binom{m}{k} (k - mx)^2 x^k (1-x)^{m-k} \]
\[ \leq \frac{\varepsilon}{2} + 2 \|f\|_{\infty} \frac{1}{4} m \frac{1}{\delta^2 m^2} < \varepsilon \]
provided \( m \) is large enough.

**Corollary B.0.3** If \( f \in C ([a,b]) \), then there exists a sequence of polynomials which converge uniformly to \( f \) on \([a,b]\).
B.1 Functions Of Many Variables

Let $f$ be a continuous function defined on $[0, 1]$ and so if $\varepsilon > 0$ is given, there exists a polynomial $p$ such that for all $x \in [0, 1]$,

$$|p(x) - f \circ l(x)| < \varepsilon$$

Therefore, letting $y = l(x)$, it follows that for all $y \in [a, b]$,

$$|p(y) - f(y)| < \varepsilon.$$

As another corollary, here is the version which will be used in Stone’s generalization later.

**Corollary B.0.4** Let $f$ be a continuous function defined on $[-M, M]$ with $f(0) = 0$. Then there exists a sequence of polynomials $\{p_n\}$ such that $p_n(0) = 0$ and $\lim_{n \to \infty} \|p_n - f\|_\infty = 0$.

**Proof:** From Corollary B.0.3 there exists a sequence of polynomials $\{\hat{p}_n\}$ such that $\|\hat{p}_n - f\|_\infty \to 0$. Simply consider $p_n = \hat{p}_n - \hat{p}_n(0)$.

### B.1 Functions Of Many Variables

Let $f$ be a continuous function defined on $[0, 1]$. Let $p_n$ be the polynomial defined by

$$p_n(x) = \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k}.$$  \hspace{1cm} (2.1)

For $f$ a continuous function defined on $[0, 1]^n$ and for $x = (x_1, \ldots, x_n)$, consider the polynomial,

$$p_m(x) = \sum_{k_1=0}^{m_1} \cdots \sum_{k_n=0}^{m_n} \binom{m_1}{k_1} \binom{m_2}{k_2} \cdots \binom{m_n}{k_n} x_1^{k_1} (1-x_1)^{m_1-k_1} x_2^{k_2} (1-x_2)^{m_2-k_2} \cdots x_n^{k_n} (1-x_n)^{m_n-k_n}.$$ \hspace{1cm} (2.2)

Also define if $I$ is a set in $\mathbb{R}^n$

$$\|h\|_I \equiv \sup \{|h(x)| : x \in I\}.$$

Let

$$\min(m) \equiv \min\{m_1, \ldots, m_n\}, \quad \max(m) \equiv \max\{m_1, \ldots, m_n\}$$

**Definition B.1.1** Define $p_m$ converges uniformly to $f$ on a set, $I$ if

$$\lim_{\min(m) \to \infty} \|p_m - f\|_I = 0.$$  

To simplify the notation, let $k = (k_1, \ldots, k_n)$ where each $k_i \in [0, m_i]$,

$$\frac{k}{m} \equiv \left(\frac{k_1}{m_1}, \ldots, \frac{k_n}{m_n}\right),$$

and let

$$\binom{m}{k} \equiv \binom{m_1}{k_1} \binom{m_2}{k_2} \cdots \binom{m_n}{k_n}.$$  

Also define for $k = (k_1, \ldots, k_n)$

$$x^k (1-x)^{m-k} \equiv x_1^{k_1} (1-x_1)^{m_1-k_1} x_2^{k_2} (1-x_2)^{m_2-k_2} \cdots x_n^{k_n} (1-x_n)^{m_n-k_n}.$$  

Thus in terms of this notation,

$$p_m(x) = \sum_{k \leq m} \binom{m}{k} x^k (1-x)^{m-k} f\left(\frac{k}{m}\right)$$

This is the $n$ dimensional version of the Bernstein polynomials which is what results in the case where $n = 1$. 

---

**Proof:** Let $l : [0, 1] \to [a, b]$ be one to one, linear and onto. Then $f \circ l$ is continuous on $[0, 1]$ and so if $\varepsilon > 0$ is given, there exists a polynomial $p$ such that for all $x \in [0, 1]$,

$$|p(x) - f \circ l(x)| < \varepsilon$$

Therefore, letting $y = l(x)$, it follows that for all $y \in [a, b]$,

$$|p(y) - f(y)| < \varepsilon.$$
Lemma B.1.2 For \( x \in [0, 1]^n \), \( f \) a continuous \( \mathbb{F} \) valued function defined on \([0, 1]^n\), and \( p_m \) given in (2.4), \( p_m \) converges uniformly to \( f \) on \([0, 1]^n\) as \( m \to \infty \). More generally, one can have \( f \) a continuous function with values in an arbitrary real or complex normed linear space. There is no change in the conclusions and proof. You just write \( \| \cdot \| \) instead of \( | \cdot | \).

**Proof:** The function \( f \) is uniformly continuous because it is continuous on a sequentially compact set \([0, 1]^n\). Therefore, there exists \( \delta > 0 \) such that if \( |x - y| < \delta \), then

\[
|f(x) - f(y)| < \varepsilon.
\]

Denote by \( G \) the set of \( k \) such that \( (k_i - m_i x_i)^2 < \eta^2 m_i^2 \) for each \( i \) where \( \eta = \delta / \sqrt{n} \). Note this condition is equivalent to saying that for each \( i \), \( \left| \frac{k_i}{m_i} - x_i \right| < \eta \) and

\[
\left| \frac{k}{m} - x \right| < \delta
\]

A short computation shows that by the binomial theorem,

\[
\sum_{k \leq m} \binom{m}{k} x^k (1 - x)^{m-k} = 1
\]

and so for \( x \in [0, 1]^n \),

\[
|p_m(x) - f(x)| \leq \sum_{k \leq m} \binom{m}{k} x^k (1 - x)^{m-k} \left| f\left( \frac{k}{m} \right) - f(x) \right|
\]

\[
\leq \sum_{k \in G} \binom{m}{k} x^k (1 - x)^{m-k} \left| f\left( \frac{k}{m} \right) - f(x) \right|
\]

\[
+ \sum_{k \in G^c} \binom{m}{k} x^k (1 - x)^{m-k} \left| f\left( \frac{k}{m} \right) - f(x) \right|
\]

(2.3)

Now for \( k \in G \) it follows that for each \( i \)

\[
\left| \frac{k_i}{m_i} - x_i \right| < \frac{\delta}{\sqrt{n}}
\]

(2.4)

and so \( |f\left( \frac{k}{m} \right) - f\left( x \right)| < \varepsilon \) because the above implies \( \left| \frac{k}{m} - x \right| < \delta \). Therefore, the first sum on the right in (2.4) is no larger than

\[
\sum_{k \in G} \binom{m}{k} x^k (1 - x)^{m-k} \varepsilon \leq \sum_{k \leq m} \binom{m}{k} x^k (1 - x)^{m-k} \varepsilon = \varepsilon
\]

Letting \( M \geq \max \{ |f(x)| : x \in [0, 1]^n \} \) it follows that for some \( j \),

\[
\left| \frac{k_j}{m_j} - x_j \right| \geq \frac{\delta}{\sqrt{n}}, (k_j - m_j x_j)^2 \geq m_j^2 \frac{\delta^2}{n}
\]

by Lemma [436].

\[
|p_m(x) - f(x)|
\]

\[
\leq \varepsilon + 2M \sum_{k \in G^c} \binom{m}{k} x^k (1 - x)^{m-k}
\]

\[
\leq \varepsilon + 2Mn \sum_{k \in G^c} \binom{m}{k} \frac{(k_j - m_j x_j)^2}{\delta^2 m_j^2} x^k (1 - x)^{m-k}
\]

\[
\leq \varepsilon + 2Mn \frac{1}{\delta^2 m_j^2} \frac{1}{4} m_j = \varepsilon + \frac{1}{2} M \frac{n}{\delta^2 m_j} \leq \varepsilon + \frac{1}{2} M \frac{n}{\delta^2 \min (m)}
\]

(2.5)
Therefore, since the right side does not depend on \( x \), it follows that if \( \min(m) \) is large enough,

\[
||p_m - f||_{[0,1]^n} \leq 2 \varepsilon
\]

and since \( \varepsilon \) is arbitrary, this shows \( \lim_{\min(m) \to \infty} ||p_m - f||_{[0,1]^n} = 0 \). This proves the lemma. \( \blacksquare \)

You don’t have to be limited to \([0,1]^n\). A similar formula will apply for \([-a, a]^n\). Also, there are some other conclusions which are easy to get about these approximations.

**Corollary B.1.3** Let \( f \) be continuous on \([-a,a]^n\) where \( f \) has values in \( F \) or more generally in some complete normed linear space. Let \( p_m(t) \) be the Bernstein polynomials for \( g(t), t \in [0,1]^n \), where \( g = f \circ r \) and

\[
r_i(t) = 2at - a
\]

Thus

\[
r(t) = \begin{pmatrix} 2at_1 - a & 2at_2 - a & \cdots & 2at_n - a \end{pmatrix}, \quad t \in [0,1]^n
\]

and

\[
t = \begin{pmatrix} \frac{x_1 + a}{2a} & \frac{x_2 + a}{2a} & \cdots & \frac{x_n + a}{2a} \end{pmatrix}
\]

Then letting \( q_m(x) = p_m \left( \frac{x_1 + a}{2a}, \frac{x_2 + a}{2a}, \cdots, \frac{x_n + a}{2a} \right) \), it follows that \( q_m(x) \) converges uniformly to \( f(x) \) on \([-a,a]^n\) and

\[
q_m(x) = \sum_{k \leq m} \binom{m}{k} t^k (1-t)^{m-k} f \left( 2a \left( \frac{k_1}{m} \right) - a, 2a \left( \frac{k_2}{m} \right) - a, \cdots, 2a \left( \frac{k_n}{m} \right) - a \right)
\]

where

\[
t_i = \frac{x_i + a}{2a}, \quad (1-t_i) = \frac{a-x_i}{2a}
\]

Thus, \( q_m \) converges uniformly to \( f \) on \([-a,a]^n\) and also if \( f \) has values in \([0,\infty)\), then each \( q_m \) also has values in \([0,\infty)\).

**Proof:** This is pretty obvious from the formula. You have

\[
|q_m(x) - f(x)| = |p_m(t) - f(r(t))|
\]

which is known to be uniformly small for \( t \in [0,1] \) thanks to the above theorem applied to \( f \circ r \). Thus,

\[
\sup_{|x| \leq a} |q_m(x) - f(x)| = \sup_{t \in [0,1]^n} |p_m(t) - f(r(t))| < \varepsilon
\]

provided \( m \) is large enough.

Now consider the terms. If \( f \) has nonnegative values, then

\[
f \left( 2a \left( \frac{k_1}{m} \right) - a, 2a \left( \frac{k_2}{m} \right) - a, \cdots, 2a \left( \frac{k_n}{m} \right) - a \right) \geq 0
\]

Also, the product making up \( t^k (1-t)^{m-k} \) consists of the product of nonnegative numbers because the \( t_i \) and \( 1-t_i \) are all nonnegative. \( \blacksquare \)

These Bernstein polynomials are very remarkable approximations. It turns out that if \( f \) is \( C^1([0,1]^n) \), then

\[
\lim_{\min(m) \to \infty} p_{mx_i}(x) \to f_{x_i}(x) \quad \text{uniformly on} \quad [0,1]^n.
\]

We show this first for the case that \( n = 1 \). From this, it is obvious for the general case.
Lemma B.1.4 Let \( f \in C^1([0,1]) \) and let
\[
p_m(x) = \sum_{k=0}^{m} \binom{m}{k} x^k (1-x)^{m-k} f \left( \frac{k}{m} \right)
\]
be the \( m^{th} \) Bernstein polynomial. Then in addition to \( \|p_m - f\|_{[0,1]} \to 0 \), it also follows that
\[
\|p'_m - f'\|_{[0,1]} \to 0
\]

**Proof:** From simple computations,
\[
p'_m(x) = \sum_{k=1}^{m} \binom{m}{k} k x^{k-1} (1-x)^{m-k} f \left( \frac{k}{m} \right)
- \sum_{k=0}^{m-1} \binom{m}{k} x^k (m-k) (1-x)^{m-1-k} f \left( \frac{k}{m} \right)
\]
\[
= \sum_{k=1}^{m} \frac{m(m-1)!}{(m-k)!(k-1)!} x^{k-1} (1-x)^{m-k} f \left( \frac{k}{m} \right)
- \sum_{k=0}^{m-1} \binom{m}{k} x^k (m-k) (1-x)^{m-1-k} f \left( \frac{k}{m} \right)
\]
\[
= \sum_{k=0}^{m-1} \frac{m(m-1)!}{(m-1-k)!(k)!} x^k (1-x)^{m-1-k} \left( f \left( \frac{k+1}{m} \right) - f \left( \frac{k}{m} \right) \right)
- \sum_{k=0}^{m-1} \frac{m(m-1)!}{(m-1-k)!(k)!} x^k (1-x)^{m-1-k} \left( f \left( \frac{k+1}{m} \right) - f \left( \frac{k}{m} \right) \right)
\]
\[
= \sum_{k=0}^{m-1} \binom{m-1}{k} x^k (1-x)^{m-1-k} \left( f \left( \frac{k+1}{m} \right) - f \left( \frac{k}{m} \right) \right)
\]

By the mean value theorem,
\[
\frac{f \left( \frac{k+1}{m} \right) - f \left( \frac{k}{m} \right)}{1/m} = f'(x_{k,m}), \quad x_{k,m} \in \left( \frac{k}{m}, \frac{k+1}{m} \right)
\]

Now the desired result follows as before from the uniform continuity of \( f' \) on \([0,1] \). Let \( \delta > 0 \) be such that if
\[
|x - y| < \delta, \text{ then } |f'(x) - f'(y)| < \varepsilon
\]
and let \( m \) be so large that \( 1/m < \delta/2 \). Then if \( |x - \frac{k}{m}| < \delta/2 \), it follows that \( |x - x_{k,m}| < \delta \) and so
\[
|f'(x) - f'(x_{k,m})| = \left| f'(x) - \frac{f \left( \frac{k+1}{m} \right) - f \left( \frac{k}{m} \right)}{1/m} \right| < \varepsilon.
\]

Now as before, letting \( M \geq |f'(x)| \) for all \( x \),
\[
|p'_m(x) - f'(x)| \leq \sum_{k=0}^{m-1} \binom{m-1}{k} x^k (1-x)^{m-1-k} |f'(x_{k,m}) - f'(x)|
\]
Let $B.1.2$ there exists a sequence of

$$\sum_{\{x| |x - \frac{m}{k}| < \frac{1}{4}\}} \binom{m - 1}{k} x^k (1 - x)^{m - 1 - k} \varepsilon + M \sum_{k=0}^{m-1} \binom{m - 1}{k} \frac{4(k - m\varepsilon)^2}{m^2} x^k (1 - x)^{m - 1 - k} \leq \varepsilon + 4M \frac{1}{4} \frac{1}{m^2} \varepsilon^2 = \varepsilon + M \frac{1}{m\varepsilon^2} < 2\varepsilon$$

whenever $m$ is large enough. Thus this proves uniform convergence. ■

Now consider the case where $n \geq 1$. Applying the same manipulations to the sum which corresponds to the $i^{th}$ variable,

$$p_{mx_i}(x) \equiv \sum_{k_1=0}^{m_1} \cdots \sum_{k_{i-1}=0}^{m_{i-1}} \sum_{k_n=0}^{m_n} \binom{m_1}{k_1} \cdots \binom{m_{i-1}}{k_{i-1}} \cdots \binom{m_n}{k_n} x_1^{k_1} (1 - x_1)^{m_1-k_1} \cdots x_i^{k_i} (1 - x_i)^{m_i-k_i} \cdots x_n^{k_n} (1 - x_n)^{m_n-k_n}.$$  

By the mean value theorem, the difference quotient is of the form

$$f_{x_i}(\mathbf{x}_{k,m}),$$

the $i^{th}$ component of $\mathbf{x}_{k,m}$ being between $\frac{k_i}{m_i}$ and $\frac{k_i+1}{m_i}$. Therefore, a repeat of the above argument involving splitting the sum into two pieces, one for which $k/m$ is close to $x$, hence close to $\mathbf{x}_{k,m}$ and one for which some $k_j/m_j$ is not close to $x_j$ for some $j$ yields the same conclusion about uniform convergence on $[0,1]^n$. This has essentially proved the following lemma.

**Lemma B.1.5** Let $f$ be in $C^k([0,1]^n)$. Then there exists a sequence of polynomials $p_m(x)$ such that each partial derivative up to order $k$ converges uniformly to the corresponding partial derivative of $f$.

**Proof:** It was shown above that letting $\mathbf{m} = (m_1, m_2, \cdots, m_n)$,

$$\lim_{\min(m) \to \infty} \|p_m - f\|_{[0,1]^n} = 0, \quad \lim_{\min(m) \to \infty} \|p_{mx_i} - f_{x_i}\|_{[0,1]^n} = 0$$

for each $x_i$. Extending to higher derivatives is just a technical generalization of what was just shown. ■

**Theorem B.1.6** Let $f$ be a continuous function defined on

$$R \equiv \prod_{k=1}^{n} [a_k, b_k].$$

Then there exists a sequence of polynomials $\{p_m\}$ converging uniformly to $f$ on $R$ as $\min(m) \to \infty$. If $f$ is $C^k(R)$, then the partial derivatives of $p_m$ up to order $k$ converge uniformly to the corresponding partial derivatives of $f$.

**Proof:** Let $g_k: [0,1] \to [a_k, b_k]$ be linear, one to one, and onto and let

$$\mathbf{x} = \mathbf{g}(\mathbf{y}) \equiv (g_1(y_1), g_2(y_2), \cdots, g_n(y_n)).$$

Thus $\mathbf{g}: [0, 1]^n \to \prod_{k=1}^{n} [a_k, b_k]$ is one to one, onto, and each component function is linear. Then $f \circ \mathbf{g}$ is a continuous function defined on $[0, 1]^n$. It follows from Lemma B.1.2 that there exists a sequence of
polynomials, \( \{p_m(y)\} \) each defined on \([0, 1]^n\) which converges uniformly to \(f \circ g\) on \([0, 1]^n\). Therefore, \(\{p_m(g^{-1}(x))\}\) converges uniformly to \(f(x)\) on \(R\). But
\[
y = (y_1, \ldots, y_n) = (g_1^{-1}(x_1), \ldots, g_n^{-1}(x_n))
\]
and each \(g_k^{-1}\) is linear. Therefore, \(\{p_m(g^{-1}(x))\}\) is a sequence of polynomials. As to the partial derivatives, it was shown above that
\[
\lim_{\min(m) \to \infty} \|Dp_m - D(f \circ g)\|_{[0,1]^n} = 0
\]
Now the chain rule implies that
\[
D(p_m \circ g^{-1})(x) = Dp_m(g^{-1}(x)) Dg^{-1}(x)
\]
Therefore, the following convergences are uniform in \(x \in R\).
\[
\lim_{\min(m) \to \infty} D(p_m \circ g^{-1})(x)
\]
\[
= \lim_{\min(m) \to \infty} Dp_m(g^{-1}(x)) Dg^{-1}(x)
\]
\[
= \lim_{\min(m) \to \infty} D(f \circ g)(g^{-1}(x)) Dg^{-1}(x)
\]
\[
= \lim_{\min(m) \to \infty} Df(g(g^{-1}(x))) Dg(g^{-1}(x)) Dg^{-1}(x)
\]
\[
= Df(x)
\]
The claim about higher order derivatives is more technical but follows in the same way. 

There is a more general version of this theorem which is easy to get. It depends on the Tietze extension theorem, a wonderful little result which is interesting for its own sake.

### B.2 Tietze Extension Theorem

This is an interesting theorem which holds in arbitrary normal topological spaces. In particular it holds in metric space and this is the context in which it will be discussed. First, here is a useful lemma.

**Lemma B.2.1** Let \(X\) be a metric space and let \(S\) be a nonempty subset of \(X\).
\[
dist(x, S) \equiv \inf \{d(x, z) : z \in S\}
\]
Then
\[
|\dist(x, S) - \dist(y, S)| \leq d(x, y)
\]

**Proof:** Say \(\dist(x, S) \geq \dist(y, S)\). Then letting \(\varepsilon > 0\) be given, there exists \(z \in S\) such that
\[
d(y, z) < \dist(y, S) + \varepsilon
\]
Then
\[
|\dist(x, S) - \dist(y, S)| = \dist(x, S) - \dist(y, S) \leq \dist(x, S) - (d(y, z) - \varepsilon)
\]
\[
\leq d(x, z) - (d(y, z) - \varepsilon) \leq d(x, y) + d(y, z) - d(y, z) + \varepsilon = d(x, y) + \varepsilon
\]
Since \(\varepsilon\) is arbitrary,
\[
|\dist(x, S) - \dist(y, S)| \leq d(x, y)
\]
It is similar if \(\dist(x, S) < \dist(y, S)\). 

Then this shows that \(x \to \dist(x, S)\) is a continuous real valued function.
Lemma B.2.2 Let $H, K$ be two nonempty disjoint closed subsets of $X$. Then there exists a continuous function, $g : X \to [-1/3, 1/3]$ such that $g(H) = -1/3$, $g(K) = 1/3$, $g(X) \subseteq [-1/3, 1/3]$.

Proof: Let
\[
f(x) = \frac{\text{dist}(x, H)}{\text{dist}(x, H) + \text{dist}(x, K)}.
\]
The denominator is never equal to zero because if $\text{dist}(x, H) = 0$, then $x \in H$ because $H$ is closed. (To see this, pick $h_k \in B(x, 1/k) \cap H$. Then $h_k \to x$ and since $H$ is closed, $x \in H$.) Similarly, if $\text{dist}(x, K) = 0$, then $x \in K$ and so the denominator is never zero as claimed. Hence $f$ is continuous and from its definition, $f = 0$ on $H$ and $f = 1$ on $K$. Now let $g(x) = \frac{2}{3} (f(x) - \frac{1}{2})$. Then $g$ has the desired properties. \[\]

Definition B.2.3 For $f : M \subseteq X \to \mathbb{R}$, define $\|f\|_M$ as $\sup \{|f(x)| : x \in M\}$. This is just notation. I am not claiming this is a norm.

Lemma B.2.4 Suppose $M$ is a closed set in $X$ and suppose $f : M \to [-1, 1]$ is continuous at every point of $M$. Then there exists a function, $g$ which is defined and continuous on all of $X$ such that $\|f - g\|_M < \frac{2}{3}$, $g(X) \subseteq [-1/3, 1/3]$.

Proof: Let $H = f^{-1}([-1, -1/3])$, $K = f^{-1}([1/3, 1])$. Thus $H$ and $K$ are disjoint closed subsets of $M$. Suppose first $H, K$ are both nonempty. Then by Lemma B.2.2 there exists $g$ such that $g$ is a continuous function defined on all of $X$ and $g(H) = -1/3$, $g(K) = 1/3$, and $g(X) \subseteq [-1/3, 1/3]$. It follows $\|f - g\|_M < 2/3$. If $H = \emptyset$, then $f$ has all its values in $[-1/3, 1/3]$ and so letting $g \equiv 1/3$, the desired condition is obtained. If $K = \emptyset$, let $g \equiv -1/3$. \[\]

Lemma B.2.5 Suppose $M$ is a closed set in $X$ and suppose $f : M \to [-1, 1]$ is continuous at every point of $M$. Then there exists a function $g$ which is defined and continuous on all of $X$ such that $g = f$ on $M$ and $g$ has its values in $[-1, 1]$.

Proof: Using Lemma B.2.4, let $g_1$ be such that $g_1(X) \subseteq [-1/3, 1/3]$ and
\[
\|f - g_1\|_M \leq \frac{2}{3}.
\]
Suppose $g_1, \ldots, g_m$ have been chosen such that $g_j(X) \subseteq [-1/3, 1/3]$ and
\[
\left\| f - \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} g_i \right\|_M \leq \left(\frac{2}{3}\right)^{m}.
\]
This has been done for $m = 1$. Then
\[
\left\| \left(\frac{3}{2}\right)^{m} \left( f - \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} g_i \right) \right\|_M \leq 1
\]
and so $\left(\frac{3}{2}\right)^{m} \left( f - \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} g_i \right)$ can play the role of $f$ in the first step of the proof. Therefore, there exists $g_{m+1}$ defined and continuous on all of $X$ such that its values are in $[-1/3, 1/3]$ and
\[
\left\| \left(\frac{3}{2}\right)^{m} \left( f - \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} g_i \right) - g_{m+1} \right\|_M \leq \frac{2}{3}.
\]
Hence
\[
\left\| \left( f - \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} g_i \right) - \left(\frac{2}{3}\right)^{m} g_{m+1} \right\|_M \leq \left(\frac{2}{3}\right)^{m+1}.
\]
It follows there exists a sequence, \( \{g_i\} \) such that each has its values in \([-1/3, 1/3]\) and for every \( m \in \mathbb{N} \) holds. Then let
\[
g(x) = \sum_{i=1}^{\infty} \left( \frac{2}{3} \right)^{i-1} g_i(x).
\]
It follows
\[
|g(x)| \leq \sum_{i=1}^{\infty} \left( \frac{2}{3} \right)^{i-1} g_i(x) \leq \sum_{i=1}^{m} \left( \frac{2}{3} \right)^{i-1} \frac{1}{3} \leq 1
\]
and
\[
\left| \left( \frac{2}{3} \right)^{i-1} g_i(x) \right| \leq \left( \frac{2}{3} \right)^{i-1} \frac{1}{3}
\]
so the Weierstrass M test applies and shows convergence is uniform. Therefore \( g \) must be continuous by Theorem B.2.5. The estimate \( \mathbb{M} \) implies \( f = g \) on \( M \).

The following is the Tietze extension theorem.

**Theorem B.2.6** Let \( M \) be a closed nonempty subset of \( X \) and let \( f : M \to [a,b] \) be continuous at every point of \( M \). Then there exists a function, \( g \) continuous on all of \( X \) which coincides with \( f \) on \( M \) such that \( g(X) \subseteq [a,b] \).

**Proof:** Let \( f_1(x) = 1 + \frac{2}{b-a} (f(x) - b) \). Then \( f_1 \) satisfies the conditions of Lemma B.2.5 and so there exists \( g_1 : X \to [-1,1] \) such that \( g \) is continuous on \( X \) and equals \( f_1 \) on \( M \). Let \( g(x) = (g_1(x) - 1) \left( \frac{b-a}{2} \right) + b \). This works.

With the Tietze extension theorem, here is a better version of the Weierstrass approximation theorem.

**Theorem B.2.7** Let \( K \) be a closed and bounded subset of \( \mathbb{R}^n \) and let \( f : K \to \mathbb{R} \) be continuous. Then there exists a sequence of polynomials \( \{p_m\} \) such that
\[
\lim_{m \to \infty} \left( \sup \{|f(x) - p_m(x)| : x \in K\} \right) = 0.
\]
In other words, the sequence of polynomials converges uniformly to \( f \) on \( K \).

**Proof:** By the Tietze extension theorem, there exists an extension of \( f \) to a continuous function \( g \) defined on all \( \mathbb{R}^n \) such that \( g = f \) on \( K \). Now since \( K \) is bounded, there exist intervals, \([a_k,b_k]\) such that
\[
K \subseteq \bigcap_{k=1}^{n} [a_k, b_k] = R
\]
Then by the Weierstrass approximation theorem, Theorem B.2.6 there exists a sequence of polynomials \( \{p_m\} \) converging uniformly to \( g \) on \( R \). Therefore, this sequence of polynomials converges uniformly to \( g = f \) on \( K \) as well. This proves the theorem.

By considering the real and imaginary parts of a function which has values in \( \mathbb{C} \) one can generalize the above theorem.

**Corollary B.2.8** Let \( K \) be a closed and bounded subset of \( \mathbb{R}^n \) and let \( f : K \to \mathbb{F} \) be continuous. Then there exists a sequence of polynomials \( \{p_m\} \) such that
\[
\lim_{m \to \infty} \left( \sup \{|f(x) - p_m(x)| : x \in K\} \right) = 0.
\]
In other words, the sequence of polynomials converges uniformly to \( f \) on \( K \).

More generally, the function \( f \) could have values in \( \mathbb{R}^n \). There is no change in the proof. You just use norm symbols rather than absolute values and nothing at all changes in the theorem where the function is defined on a rectangle. Then you apply the Tietze extension theorem to each component in the case the function has values in \( \mathbb{R}^n \).
Appendix C

The Hausdorff Maximal Theorem

The purpose of this appendix is to prove the equivalence between the axiom of choice, the Hausdorff maximal theorem, and the well-ordering principle. The Hausdorff maximal theorem and the well-ordering principle are very useful but a little hard to believe; so, it may be surprising that they are equivalent to the axiom of choice. First it is shown that the axiom of choice implies the Hausdorff maximal theorem, a remarkable theorem about partially ordered sets.

A nonempty set is partially ordered if there exists a partial order, \( \prec \), satisfying

\[ x \prec x \]

and

\[ \text{if } x \prec y \text{ and } y \prec z \text{ then } x \prec z. \]

An example of a partially ordered set is the set of all subsets of a given set and \( \subseteq \). Note that two elements in a partially ordered set may not be related. In other words, just because \( x, y \) are in the partially ordered set, it does not follow that either \( x \prec y \) or \( y \prec x \). A subset of a partially ordered set, \( C \), is called a chain if \( x, y \in C \) implies that either \( x \prec y \) or \( y \prec x \). If either \( x \prec y \) or \( y \prec x \) then \( x \) and \( y \) are described as being comparable. A chain is also called a totally ordered set. \( C \) is a maximal chain if whenever \( C \subseteq C' \) is a chain containing \( C \), it follows the two chains are equal. In other words \( C \) is a maximal chain if there is no strictly larger chain.

**Lemma C.0.1** Let \( F \) be a nonempty partially ordered set with partial order \( \prec \). Then assuming the axiom of choice, there exists a maximal chain in \( F \).

**Proof:** Let \( \mathcal{X} \) be the set of all chains from \( F \). For \( C \in \mathcal{X} \), let

\[ S_C = \{ x \in F \text{ such that } C \cup \{x\} \text{ is a chain strictly larger than } C \}. \]

If \( S_C = \emptyset \) for any \( C \), then \( C \) is maximal. Thus, assume \( S_C \neq \emptyset \) for all \( C \in \mathcal{X} \). Let \( f(C) \in S_C \). (This is where the axiom of choice is being used.) Let

\[ g(C) = C \cup \{ f(C) \}. \]

Thus \( g(C) \supseteq C \) and \( g(C) \setminus C = \{ f(C) \} \) = \{a single element of \( F \} \). A subset \( \mathcal{T} \) of \( \mathcal{X} \) is called a tower if

\[ \emptyset \in \mathcal{T}, \]

\[ C \in \mathcal{T} \text{ implies } g(C) \in \mathcal{T}, \]

and if \( S \subseteq \mathcal{T} \) is totally ordered with respect to set inclusion, then

\[ \cup S \in \mathcal{T}. \]

Here \( S \) is a chain with respect to set inclusion whose elements are chains.
Note that $X$ is a tower. Let $\mathcal{T}_0$ be the intersection of all towers. Thus, $\mathcal{T}_0$ is a tower, the smallest tower. Are any two sets in $\mathcal{T}_0$ comparable in the sense of set inclusion so that $\mathcal{T}_0$ is actually a chain? Let $C_0$ be a set of $\mathcal{T}_0$ which is comparable to every set of $\mathcal{T}_0$. Such sets exist, $\emptyset$ being an example. Let

$$B = \{ D \in \mathcal{T}_0 : D \supseteq C_0 \text{ and } f(C_0) \notin D \}.$$  

The picture represents sets of $B$. As illustrated in the picture, $D$ is a set of $B$ when $D$ is larger than $C_0$ but fails to be comparable to $g(C_0)$. Thus there would be more than one chain ascending from $C_0$ if $B \neq \emptyset$, rather like a tree growing upward in more than one direction from a fork in the trunk. It will be shown this can’t take place for any such $C_0$ by showing $B = \emptyset$.

![Diagram](https://via.placeholder.com/150)

This will be accomplished by showing $\mathcal{T}_0 \equiv \mathcal{T}_0 \setminus B$ is a tower. Since $\mathcal{T}_0$ is the smallest tower, this will require that $\mathcal{T}_0 = \mathcal{T}_0$ and so $B = \emptyset$.

**Claim:** $\mathcal{T}_0$ is a tower and so $B = \emptyset$.

**Proof of the claim:** It is clear that $\emptyset \in \mathcal{T}_0$ because for $\emptyset$ to be contained in $B$ it would be required to properly contain $C_0$ which is not possible. Suppose $D \in \mathcal{T}_0$. The plan is to verify $g(D) \in \mathcal{T}_0$.

**Case 1:** $f(D) \in C_0$. If $D \subseteq C_0$, then since both $D$ and $\{ f(D) \}$ are contained in $C_0$, it follows $g(D) \subseteq C_0$ and so $g(D) \notin B$. On the other hand, if $D \supseteq C_0$, then since $D \in \mathcal{T}_0$, $f(C_0) \in D$ and so $g(D)$ also contains $f(C_0)$ implying $g(D) \notin B$. These are the only two cases to consider because $C_0$ is comparable to every set of $\mathcal{T}_0$.

**Case 2:** $f(D) \notin C_0$. If $D \subseteq C_0$ it can’t be the case that $f(D) \notin C_0$ because if this were so, $g(D)$ would not compare to $C_0$.

![Diagram](https://via.placeholder.com/150)

Hence if $f(D) \notin C_0$, then $D \supseteq C_0$. If $D = C_0$, then $f(D) = f(C_0) \in g(D)$ so $g(D) \notin B$. Therefore, assume $D \supseteq C_0$. Then, since $D$ is in $\mathcal{T}_0$, $f(C_0) \in D$ and so $f(C_0) \in g(D)$. Therefore, $g(D) \in \mathcal{T}_0$.

Now suppose $S$ is a totally ordered subset of $\mathcal{T}_0$ with respect to set inclusion. Then if every element of $S$ is contained in $C_0$, so is $\cup S$ and so $\cup S \in \mathcal{T}_0$. If, on the other hand, some chain from $S$, $C$, contains $C_0$ properly, then since $C \notin B$, $f(C_0) \in C \subseteq \cup S$ showing that $\cup S \notin B$ also. This has proved $\mathcal{T}_0$ is a tower and since $\mathcal{T}_0$ is the smallest tower, it follows $\mathcal{T}_0 = \mathcal{T}_0$. This has shown roughly that no splitting into more than one ascending chain can occur at any $C_0$ which is comparable to every set of $\mathcal{T}_0$. Next it is shown that every element of $\mathcal{T}_0$ has the property that it is comparable to all other elements of $\mathcal{T}_0$. This is done by showing that these elements which possess this property form a tower.

Define $\mathcal{T}_1$ to be the set of all elements of $\mathcal{T}_0$ which are comparable to every element of $\mathcal{T}_0$. (Recall the elements of $\mathcal{T}_0$ are chains from the original partial order.)

**Claim:** $\mathcal{T}_1$ is a tower.

**Proof of the claim:** It is clear that $\emptyset \in \mathcal{T}_1$ because $\emptyset$ is a subset of every set. Suppose $C_0 \in \mathcal{T}_1$. It is necessary to verify that $g(C_0) \in \mathcal{T}_1$. Let $D \in \mathcal{T}_0$ (Thus $D \subseteq C_0$ or else $D \supseteq C_0$) and consider $g(C_0) = C_0 \cup \{ f(C_0) \}$. If $D \subseteq C_0$, then $D \subseteq g(C_0)$ so $g(C_0)$ is comparable to $D$. If $D \supseteq C_0$, then $D \supseteq g(C_0)$ by what was just shown ($B = \emptyset$). Hence $g(C_0)$ is comparable to $D$. Since $D$ was arbitrary, it follows $g(C_0)$ is comparable to every set of $\mathcal{T}_0$. Now suppose $S$ is a chain of elements of $\mathcal{T}_1$ and let $D$ be an element of $\mathcal{T}_0$. If every element in the chain, $S$ is contained in $D$, then $\cup S$ is also contained
in $\mathcal{D}$. On the other hand, if some set, $\mathcal{C}$, from $\mathcal{S}$ contains $\mathcal{D}$ properly, then $\cup \mathcal{S}$ also contains $\mathcal{D}$. Thus $\cup \mathcal{S} \in \mathcal{T}_1$ since it is comparable to every $\mathcal{D} \in \mathcal{T}_0$.

This shows $\mathcal{T}_1$ is a tower and proves therefore, that $\mathcal{T}_0 = \mathcal{T}_1$. Thus every set of $\mathcal{T}_0$ compares with every other set of $\mathcal{T}_0$ showing $\mathcal{T}_0$ is a chain in addition to being a tower.

Now $\cup \mathcal{T}_0, g(\cup \mathcal{T}_0) \in \mathcal{T}_0$. Hence, because $g(\cup \mathcal{T}_0)$ is an element of $\mathcal{T}_0$, and $\mathcal{T}_0$ is a chain of these, it follows $g(\cup \mathcal{T}_0) \subseteq \cup \mathcal{T}_0$.

a contradiction. Hence there must exist a maximal chain after all. This proves the lemma.

If $X$ is a nonempty set, $\leq$ is an order on $X$ if

$$x \leq x,$$

and if $x$, $y \in X$, then

either $x \leq y$ or $y \leq x$

and

if $x \leq y$ and $y \leq z$ then $x \leq z$.

$\leq$ is a well order and say that $(X, \leq)$ is a well-ordered set if every nonempty subset of $X$ has a smallest element. More precisely, if $S \neq \emptyset$ and $S \subseteq X$ then there exists an $x \in S$ such that $x \leq y$ for all $y \in S$. A familiar example of a well-ordered set is the natural numbers.

Lemma C.0.2 The Hausdorff maximal principle implies every nonempty set can be well-ordered.

Proof: Let $X$ be a nonempty set and let $a \in X$. Then $\{a\}$ is a well-ordered subset of $X$. Let $\mathcal{F} = \{S \subseteq X : \text{there exists a well order for } S\}$.

Thus $\mathcal{F} \neq \emptyset$. For $S_1$, $S_2 \in \mathcal{F}$, define $S_1 \prec S_2$ if $S_1 \subseteq S_2$ and there exists a well order for $S_2$, $\leq_2$ such that

$(S_2, \leq_2)$ is well-ordered

and if

$y \in S_2 \setminus S_1$ then $x \leq_2 y$ for all $x \in S_1$,

and if $\leq_1$ is the well order of $S_1$ then the two orders are consistent on $S_1$. Then observe that $\prec$ is a partial order on $\mathcal{F}$. By the Hausdorff maximal principle, let $\mathcal{C}$ be a maximal chain in $\mathcal{F}$ and let $X_\infty \equiv \cup \mathcal{C}$.

Define an order, $\leq$, on $X_\infty$ as follows. If $x$, $y$ are elements of $X_\infty$, pick $S \in \mathcal{C}$ such that $x$, $y$ are both in $S$. Then if $\leq_S$ is the order on $S$, let $x \leq y$ if and only if $x \leq_S y$. This definition is well defined because of the definition of the order, $\prec$. Now let $U$ be any nonempty subset of $X_\infty$. Then $S \cap U \neq \emptyset$ for some $S \in \mathcal{C}$. Because of the definition of $\leq$, if $y \in S_2 \setminus S_1$, $S_1 \subseteq S_2$, then $x \leq y$ for all $x \in S_1$. Thus, if $y \in X_\infty \setminus S$ then $x \leq y$ for all $x \in S$ and so the smallest element of $S \cap U$ exists and is the smallest element in $U$. Therefore $X_\infty$ is well-ordered. Now suppose there exists $z \in X \setminus X_\infty$. Define the following order, $\leq_1$, on $X_\infty \cup \{z\}$.

\[x \leq_1 y \text{ if and only if } x \leq y \text{ whenever } x, y \in X_\infty\]

\[x \leq_1 z \text{ whenever } x \in X_\infty.\]

Then let

$\tilde{\mathcal{C}} = \{S \in \mathcal{C} \text{ or } X_\infty \cup \{z\}\}.$

Then $\tilde{\mathcal{C}}$ is a strictly larger chain than $\mathcal{C}$ contradicting maximality of $\mathcal{C}$. Thus $X \setminus X_\infty = \emptyset$ and this shows $X$ is well-ordered by $\leq$. This proves the lemma.

With these two lemmas the main result follows.
Theorem C.0.3  The following are equivalent.

The axiom of choice

The Hausdorff maximal principle

The well-ordering principle.

Proof: It only remains to prove that the well-ordering principle implies the axiom of choice. Let $I$ be a nonempty set and let $X_i$ be a nonempty set for each $i \in I$. Let $X = \bigcup \{X_i : i \in I\}$ and well order $X$. Let $f(i)$ be the smallest element of $X_i$. Then

$$f \in \prod_{i \in I} X_i.$$ 

C.1  The Hamel Basis

A Hamel basis is nothing more than the correct generalization of the notion of a basis for a finite dimensional vector space to vector spaces which are possibly not of finite dimension.

Definition C.1.1  Let $X$ be a vector space. A Hamel basis is a subset of $X$, $\Lambda$ such that every vector of $X$ can be written as a finite linear combination of vectors of $\Lambda$ and the vectors of $\Lambda$ are linearly independent in the sense that if $\{x_1, \ldots, x_n\} \subseteq \Lambda$ and

$$\sum_{k=1}^{n} c_k x_k = 0$$

then each $c_k = 0$.

The main result is the following theorem.

Theorem C.1.2  Let $X$ be a nonzero vector space. Then it has a Hamel basis.

Proof: Let $x_1 \in X$ and $x_1 \neq 0$. Let $\mathcal{F}$ denote the collection of subsets of $X$, $\Lambda$ containing $x_1$ with the property that the vectors of $\Lambda$ are linearly independent as described in Definition C.1.1 partially ordered by set inclusion. By the Hausdorff maximal theorem, there exists a maximal chain, $\mathcal{C}$. Let $\Lambda = \cup \mathcal{C}$. Since $\mathcal{C}$ is a chain, it follows that if $\{x_1, \ldots, x_n\} \subseteq \mathcal{C}$ then there exists a single $\Lambda' \in \mathcal{C}$ containing all these vectors. Therefore, if

$$\sum_{k=1}^{n} c_k x_k = 0$$

it follows each $c_k = 0$. Thus the vectors of $\Lambda$ are linearly independent. Is every vector of $X$ a finite linear combination of vectors of $\Lambda$?

Suppose not. Then there exists $z$ which is not equal to a finite linear combination of vectors of $\Lambda$. Consider $\Lambda \cup \{z\}$. If

$$cz + \sum_{k=1}^{m} c_k x_k = 0$$

where the $x_i$ are vectors of $\Lambda$, then if $c \neq 0$ this contradicts the condition that $z$ is not a finite linear combination of vectors of $\Lambda$. Therefore, $c = 0$ and now all the $c_k$ must equal zero because it was just shown $\Lambda$ is linearly independent. It follows $\mathcal{C} \cup \{\Lambda \cup \{z\}\}$ is a strictly larger chain than $\mathcal{C}$ and this is a contradiction. Therefore, $\Lambda$ is a Hamel basis as claimed. This proves the theorem.
C.2 Exercises

1. Zorn’s lemma states that in a nonempty partially ordered set, if every chain has an upper bound, there exists a maximal element, $x$ in the partially ordered set. $x$ is maximal, means that if $x \prec y$, it follows $y = x$. Show Zorn’s lemma is equivalent to the Hausdorff maximal theorem.

2. Show that if $Y, Y_1$ are two Hamel bases of $X$, then there exists a one to one and onto map from $Y$ to $Y_1$. Thus any two Hamel bases are of the same size.
Bibliography


Index

$(\infty, \infty)$, 104
$C^1$, 114
$C^k$, 114
$C^1$, 114
$C^1$ and differentiability, 114
$C_c(\Omega)$, 114
$C_c^\infty$, 114
$C_c^{\ast}$, 114
$E_c$, 114
$G_\delta$, 114
$L^1$
  complex vector space, 114
$L^1_{loc}$, 114
$L^p$
  compactness, 114
  completeness, 114
  continuity of translation, 114
  definition, 114
  density of continuous functions, 114
  density of simple functions, 114
  norm, 114
  separability, 114
$L^p$
  separable, 114
$L^1(\Omega)$, 114
$L^\infty$, 114
$L^p$
  density of smooth functions, 114
$\cap$, 114
$\cup$, 114
$\epsilon$ net, 114
$\sigma$ algebra, 114

A close to B
  eigenvalues, 114
  Abel's formula, 114
  Abelian group, 114
  absolute convergence
    convergence, 114
  absolute value
    complex number, 114
  accumulation point, 114
  adjoint, 114
  adjugate, 114
  algebraic number
  minimum polynomial, 114
  algebraic numbers, 114
  field, 114
  analytic function of matrix, 114
  ann(m), 114
  approximate identity, 114
  arcwise connected, 114
  area
    parallelogram, 114
    area of a parallelogram, 114
    arithmetic mean, 114
    associated, 114
    associates, 114
    at most countable, 114
    augmented matrix, 114
    axiom of choice, 114
    back substitution, 114
    balls
      disjoint, almost covering a set, 114
      Banach space, 114
    basic variables, 114
    basis, 114
    basis of eigenvectors
      diagonalizable, 114
    basis of vector space, 114
    Bernstein polynomial
      approximation of derivative, 114
    Bernstein polynomials, 114
    Besicovitch
      covering theorem, 114
    Besicovitch covering theorem, 114
    Binet Cauchy
      volumes, 114
    Binet Cauchy formula, 114
    binomial theorem, 114
    block diagonal matrix, 114
    block matrix, 114
    block multiplication, 114
    Borel sets, 114
    bounded continuous linear functions, 114
    bounded linear transformations, 114
    box product, 114

Cantar function, 114

611
Cantor set, 484
Caratheodory’s procedure, 478
Cauchy Schwarz inequality, 254
Cauchy sequence, 555, 305
Cayley Hamilton theorem, 177, 182, 183, 212, 349
chain rule, 407
change of variables, 524
better, 528
formula, 526
linear map, 519
map not one to one, 529
change of variables general case, 529
closed
connected component, 491
connected components, 491
equivalence class, 491
equivalence relation, 491
open sets, 491
connected sets
intersection, 491
intervals, 491
real line, 491
consistent, 405
continuous function, 483
continuous functions, 481
equivalent conditions, 483
contraction map, 248
fixed point, 248
fixed point theorem, 248
convergence in measure, 494
convex
set, 494
functions, 494
convex combination, 115, 268
convex hull, 115, 268
compactness, 115
convex
set, 494
functions, 494
convex combination, 115, 268
convex hull, 115, 268
compactness, 115
convolution, 545
Coordinates, 54
countable, 4
counting zeros, 315
Courant Fischer theorem, 336
Cramer's rule, 174
cross product, 587
area of parallelogram, 590
distributive law, 591
geometric description, 591
cross product
coordinate description, 587
geometric description, 587
parallelepiped, 590
cyclic decomposition theorem, 494
cyclic set, 494
independence, 494
De Moivre’s theorem, 11
definition of $L^p$, 536
definition of a $C^k$ function, 414
density of continuous functions in $L^p$, 542
derivative, 407
difference, 407
chain rule, 407
continuity, 407
continuity of Gateaux derivative, 410
continuous, 407
continuous Gateaux derivatives, 410
Frechet, 407
INDEX

Gateaux, 408

generalized partial, 410

higher order, 413

matrix, 408

partial, 416

second, 413

well defined, 406

derivatives, 406

determinant

definition, 167

estimate for Hermitian matrix, 352

expansion along row, column, 172

Hadamard inequality, 352

matrix inverse, 173

partial derivative, cofactor, 181

permutation of rows, 167

product, 169

product of eigenvalues, 319

row, column operations, 169

summary of properties, 176

symmetric definition, 168

transpose, 168

diagonal matrix, 132

diagonalizability, 132

diagonalizable, 132

formal derivative, 136

minimal polynomial and its derivative, 136

differentiable, 406

continuous, 406

continuous partials, 417

differentiable function

measurable sets, 523

sets of measure zero, 523

differential equations

first order systems, 322

dimension of a vector space, 252

dimension of vector space, 55

direct sum, 117

torsion module, 193

directional derivative, 410

discrete Fourier transform, 349

distance, 237

distance to a nonempty set, 239

distinct eigenvalues, 142

dominated convergence

generalization, 135

dominated convergence theorem, 135

dot product, 253

dyadics, 103

Dynkin’s lemma, 467

echelon form, 38

Eggoroff theorem, 533

eigen-pair, 129

eigenvalue, 129

existence, 130

eigenvalues, 129

AB and BA, 130

eigenvector, 129

existence, 130

eigenvectors

distinct eigenvalues, 142

independent, 142

elementary matrices, 142

elementary matrix

inverse, 142

properties, 142

elementary operations, 142

elementary symmetric polynomials, 216

empty set, 2

equality of mixed partial derivatives, 417

equivalence class, 5

equivalence relation, 5

Euclidean algorithm, 21

Euclidean domain, 185

existence of a fixed point, 367

factorization of a matrix

general p.i.d., 197

factorization of matrix

Euclidean domain, 197

Fatou’s lemma, 498

field axioms, 8

field extension, 65

dimension, 66

finite, 66

field extensions, 67

Field of scalars, 65

finite dimensional vector space, 67

finite measure

regularity, 469

flip, 442

formal derivative, 136

Fourier series, 304

Frechet derivative, 406

Fredholm alternative, 300

free variables, 44

Frobenius

inner product, 324

Frobenius norm, 262

— singular value decomposition, 341

Frobenius norm, 341

Fubini theorem, 512

functions

measurable, 512

fundamental theorem of algebra, 12

613
<table>
<thead>
<tr>
<th>Term</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>fundamental theorem of algebra</td>
<td></td>
</tr>
<tr>
<td>plausibility argument,</td>
<td>14</td>
</tr>
<tr>
<td>rigorous proof,</td>
<td>15</td>
</tr>
<tr>
<td>fundamental theorem of arithmetic,</td>
<td>23</td>
</tr>
<tr>
<td>fundamental theorem of calculus</td>
<td></td>
</tr>
<tr>
<td>general Radon measures,</td>
<td>548</td>
</tr>
<tr>
<td>Radon measures,</td>
<td>549</td>
</tr>
<tr>
<td>g.c.d.,</td>
<td>17</td>
</tr>
<tr>
<td>Gamma function,</td>
<td>551</td>
</tr>
<tr>
<td>gamma function,</td>
<td>423</td>
</tr>
<tr>
<td>Gateaux derivative,</td>
<td></td>
</tr>
<tr>
<td>continuous,</td>
<td>414</td>
</tr>
<tr>
<td>Gauss Elimination,</td>
<td>46</td>
</tr>
<tr>
<td>Gauss elimination,</td>
<td>37</td>
</tr>
<tr>
<td>Gauss Jordan method for inverses,</td>
<td>84</td>
</tr>
<tr>
<td>Gauss Seidel method,</td>
<td>363</td>
</tr>
<tr>
<td>Gelfand,</td>
<td>506</td>
</tr>
<tr>
<td>generalized eigenspaces,</td>
<td>150</td>
</tr>
<tr>
<td>generalized eigenvectors,</td>
<td>151</td>
</tr>
<tr>
<td>geometric mean,</td>
<td>459</td>
</tr>
<tr>
<td>Gerschgorin’s theorem,</td>
<td>314</td>
</tr>
<tr>
<td>Gram Schmidt process,</td>
<td>259</td>
</tr>
<tr>
<td>greatest common divisor,</td>
<td>18</td>
</tr>
<tr>
<td>characterization,</td>
<td>24</td>
</tr>
<tr>
<td>Gronwall’s inequality,</td>
<td>372</td>
</tr>
<tr>
<td>Hadamard</td>
<td></td>
</tr>
<tr>
<td>inequality,</td>
<td>552</td>
</tr>
<tr>
<td>Hahn decomposition,</td>
<td>588</td>
</tr>
<tr>
<td>Hahn Jordan decomposition,</td>
<td>588</td>
</tr>
<tr>
<td>Hamel basis,</td>
<td>241</td>
</tr>
<tr>
<td>Hardy’s inequality,</td>
<td>531</td>
</tr>
<tr>
<td>Hausdorff maximal theorem,</td>
<td>536</td>
</tr>
<tr>
<td>Hermitian</td>
<td></td>
</tr>
<tr>
<td>orthonormal basis eigenvectors,</td>
<td>633</td>
</tr>
<tr>
<td>positive definite,</td>
<td>635</td>
</tr>
<tr>
<td>Hermitian matrix</td>
<td></td>
</tr>
<tr>
<td>factorization,</td>
<td>652</td>
</tr>
<tr>
<td>positive part,</td>
<td>531</td>
</tr>
<tr>
<td>positive part, Lipschitz continuous,</td>
<td>537</td>
</tr>
<tr>
<td>Hermitian operator,</td>
<td>467</td>
</tr>
<tr>
<td>largest, smallest, eigenvalues,</td>
<td>630</td>
</tr>
<tr>
<td>Hessian matrix</td>
<td>467</td>
</tr>
<tr>
<td>higher order derivative</td>
<td></td>
</tr>
<tr>
<td>multilinear form,</td>
<td>413</td>
</tr>
<tr>
<td>higher order derivatives,</td>
<td>430</td>
</tr>
<tr>
<td>implicit function theorem,</td>
<td>531</td>
</tr>
<tr>
<td>inverse function theorem,</td>
<td>531</td>
</tr>
<tr>
<td>Hilbert space</td>
<td>555</td>
</tr>
<tr>
<td>Holder’s inequality,</td>
<td>256</td>
</tr>
<tr>
<td>functions in $L^1$,</td>
<td>494</td>
</tr>
<tr>
<td>linear,</td>
<td>494</td>
</tr>
<tr>
<td>operator valued function,</td>
<td>572</td>
</tr>
<tr>
<td>vector valued function,</td>
<td>572</td>
</tr>
<tr>
<td>integral</td>
<td></td>
</tr>
<tr>
<td>decreasing function,</td>
<td>587</td>
</tr>
<tr>
<td>functions in $L^1$,</td>
<td>587</td>
</tr>
<tr>
<td>linear,</td>
<td>587</td>
</tr>
<tr>
<td>operator valued function,</td>
<td>572</td>
</tr>
<tr>
<td>vector valued function,</td>
<td>572</td>
</tr>
<tr>
<td>integral domain</td>
<td></td>
</tr>
<tr>
<td>prime elements,</td>
<td>189</td>
</tr>
<tr>
<td>integral over a measurable set,</td>
<td>585</td>
</tr>
<tr>
<td>interchange order of integration,</td>
<td>524</td>
</tr>
<tr>
<td>interior point</td>
<td>467</td>
</tr>
<tr>
<td>intermediate value theorem,</td>
<td>467</td>
</tr>
<tr>
<td>intersection</td>
<td></td>
</tr>
<tr>
<td>intervals</td>
<td></td>
</tr>
<tr>
<td>notation,</td>
<td>11</td>
</tr>
<tr>
<td>invariance of domain,</td>
<td>551</td>
</tr>
<tr>
<td>invariant subspaces</td>
<td></td>
</tr>
<tr>
<td>direct sum, block diagonal matrix,</td>
<td>585</td>
</tr>
<tr>
<td>inverse function theorem,</td>
<td>585</td>
</tr>
<tr>
<td>higher order derivatives,</td>
<td>585</td>
</tr>
<tr>
<td>inverse image</td>
<td>6</td>
</tr>
<tr>
<td>inverses and determinants,</td>
<td>178</td>
</tr>
<tr>
<td>invertible,</td>
<td>58</td>
</tr>
<tr>
<td>invertible maps</td>
<td>128</td>
</tr>
<tr>
<td>different spaces,</td>
<td>128</td>
</tr>
<tr>
<td>irreducible,</td>
<td>18</td>
</tr>
<tr>
<td>relatively prime,</td>
<td>18</td>
</tr>
<tr>
<td>isomorphism</td>
<td>128</td>
</tr>
<tr>
<td>iterative methods</td>
<td></td>
</tr>
<tr>
<td>alternate proof of convergence,</td>
<td>571</td>
</tr>
<tr>
<td>convergence criterion,</td>
<td>571</td>
</tr>
<tr>
<td>diagonally dominant,</td>
<td>571</td>
</tr>
<tr>
<td>proof of convergence,</td>
<td>571</td>
</tr>
<tr>
<td>Jensens inequality</td>
<td>571</td>
</tr>
<tr>
<td>Jocobi method</td>
<td>361</td>
</tr>
<tr>
<td>Jordan block</td>
<td>150</td>
</tr>
<tr>
<td>Jordan canonical form,</td>
<td>150</td>
</tr>
<tr>
<td>existence and uniqueness,</td>
<td>150</td>
</tr>
<tr>
<td>powers of a matrix,</td>
<td>150</td>
</tr>
<tr>
<td>ideal,</td>
<td>185</td>
</tr>
<tr>
<td>maximal,</td>
<td>185</td>
</tr>
<tr>
<td>principal,</td>
<td>185</td>
</tr>
<tr>
<td>higher function theorem,</td>
<td>154</td>
</tr>
<tr>
<td>higher order derivatives,</td>
<td>154</td>
</tr>
<tr>
<td>inconsistent,</td>
<td>154</td>
</tr>
<tr>
<td>initial value problem</td>
<td>154</td>
</tr>
<tr>
<td>uniqueness,</td>
<td>154</td>
</tr>
<tr>
<td>inner product</td>
<td>468</td>
</tr>
<tr>
<td>inner product space</td>
<td>555</td>
</tr>
<tr>
<td>adjoint operator,</td>
<td>486</td>
</tr>
<tr>
<td>compact sets,</td>
<td>105</td>
</tr>
<tr>
<td>inner regular,</td>
<td>105</td>
</tr>
<tr>
<td>inner regularity,</td>
<td>105</td>
</tr>
<tr>
<td>integral</td>
<td></td>
</tr>
<tr>
<td>decreasing function,</td>
<td>105</td>
</tr>
<tr>
<td>functions in $L^1$,</td>
<td>105</td>
</tr>
<tr>
<td>linear,</td>
<td>105</td>
</tr>
<tr>
<td>operator valued function,</td>
<td>572</td>
</tr>
<tr>
<td>vector valued function,</td>
<td>572</td>
</tr>
<tr>
<td>integral domain</td>
<td></td>
</tr>
<tr>
<td>prime elements,</td>
<td>105</td>
</tr>
<tr>
<td>prime elements,</td>
<td>105</td>
</tr>
<tr>
<td>integral over a measurable set,</td>
<td>105</td>
</tr>
<tr>
<td>interchange order of integration,</td>
<td>105</td>
</tr>
<tr>
<td>interior point</td>
<td>105</td>
</tr>
<tr>
<td>intermediate value theorem,</td>
<td>105</td>
</tr>
<tr>
<td>intersection</td>
<td>105</td>
</tr>
<tr>
<td>intervals</td>
<td></td>
</tr>
<tr>
<td>notation,</td>
<td>105</td>
</tr>
<tr>
<td>invariance of domain,</td>
<td>105</td>
</tr>
<tr>
<td>invariant subspaces</td>
<td></td>
</tr>
<tr>
<td>direct sum, block diagonal matrix,</td>
<td>105</td>
</tr>
<tr>
<td>inverse function theorem,</td>
<td>105</td>
</tr>
<tr>
<td>higher order derivatives,</td>
<td>105</td>
</tr>
<tr>
<td>inverse image</td>
<td>105</td>
</tr>
<tr>
<td>inverses and determinants,</td>
<td>105</td>
</tr>
<tr>
<td>invertible,</td>
<td>105</td>
</tr>
<tr>
<td>invertible maps</td>
<td>105</td>
</tr>
<tr>
<td>different spaces,</td>
<td>105</td>
</tr>
<tr>
<td>irreducible,</td>
<td>105</td>
</tr>
<tr>
<td>relatively prime,</td>
<td>105</td>
</tr>
<tr>
<td>isomorphism</td>
<td>105</td>
</tr>
<tr>
<td>iterative methods</td>
<td></td>
</tr>
<tr>
<td>alternate proof of convergence,</td>
<td>105</td>
</tr>
<tr>
<td>convergence criterion,</td>
<td>105</td>
</tr>
<tr>
<td>diagonally dominant,</td>
<td>105</td>
</tr>
<tr>
<td>proof of convergence,</td>
<td>105</td>
</tr>
<tr>
<td>Jensens inequality</td>
<td>105</td>
</tr>
<tr>
<td>Jocobi method</td>
<td>105</td>
</tr>
<tr>
<td>Jordan block</td>
<td>105</td>
</tr>
<tr>
<td>Jordan canonical form,</td>
<td>105</td>
</tr>
<tr>
<td>existence and uniqueness,</td>
<td>105</td>
</tr>
<tr>
<td>powers of a matrix,</td>
<td>105</td>
</tr>
</tbody>
</table>
INDEX

kernel of a product, 126
  direct sum decomposition, 126
Kirchhoff’s law, 49, 71
Krylov sequence, 143
Lagrange multipliers, 434, 435
Laplace expansion, 171
leading entry, 38
least squares, 406
least squares regression, 122
Lebesgue decomposition, 558
Lebesgue integral
desires to be linear, 482
nonnegative function, 488
other definitions, 491
simple function, 489
Lebesgue measurable but not Borel, 487
Lebesgue measure
one dimensional, 482
translation invariance, 482
translation invariant, 513
Lebesgue Stieltjes measure, 481
lim inf, 25
  properties, 27
lim sup, 25
  properties, 27
lim sup and lim inf
  limit, 257
  limit point, 258
limit
  continuity, 406
  infinite limits, 211
  limit of a function, 406
  limit of a sequence, 406
  well defined, 206
  limit point, 206
limits
  combinations of functions, 206
  existence of limits, 406
  limits and continuity, 406
Lindeloff property, 244
Lindemann Weierstrass theorem, 218
Lindemann Weierstrass theorem, 224
linear combination, 53, 169
linear independence, 253
linear maps
  continuous, 206
  equivalent conditions, 206
linear transformation, 53
  defined on a basis, 132
  dimension of vector space, 118
  kernel, 122
  rank m, 138
linear transformations
  a vector space, 118
  commuting, 123
  composition, matrices, 111
  sum, 110
  linearly independent, 53
  linearly independent set
    enlarging to a basis, 496
  Lipschitz continuous, 411
  measurable sets, 418
  sets of measure zero, 418
  little o notation, 110
  local maximum, 110
  local minimum, 110
  locally one to one, 110
Lusin, 533
Lusin’s theorem, 550
map
  $C^1$, 118
  primitive and flips, 118
Markov matrix, 273
  limit, 276
  regular, 276
  steady state, 276
math induction, 6
  mathematical induction, 6
matrices
  commuting, 206
  notation, 76
  transpose, 84
matrix
  differentiation operator, 118
  inverse, 83
  left inverse, 173
  linear transformation, 118
  lower triangular, 174
  main diagonal, 175
  Markov, 273
  polynomial, 182
  right inverse, 173
  right, left inverse, 173
  row, column, determinant rank, 173
  stochastic, 273
  upper triangular, 173
matrix
  positive definite, 351
matrix exponential, 371
matrix multiplication
  definition, 76
  entries of the product, 76
  not commutative, 76
  properties, 76
  vectors, 76
maximal function
Radon measures, 547
maximal ideal, 213
mean value inequality, 409, 428
mean value theorem, 409, 428
measurability
  limit of simple functions, 462
measurable, 461
  complex valued, 461
  equivalent formulations, 462
  linear combinations, 462
measurable complex functions
  simple functions, 462
measurable function
  pointwise a.e. limit of continuous, 510
  a.e. Borel, 511
measurable functions, 461
  approximation, 461
  pointwise limit, 462
  simple functions, 462
measurable into \((0, 1]\), 546
measurable sets, 474
measure, 470
  \(\sigma\) finite, 471
  Borel, 471
  inner regular, 475
  metric space, 475
  outer regular, 475
  properties, 475
measure of balls, 515
measure space
  completion, 481
  regular, 481
measures
  Lebesgue decomposition, 559
  absolutely continuous, 559
  decreasing sequences of sets, 559
  increasing sequences of sets, 559
  regularity, 559
measures from outer measures, 478
metric, 256
  properties, 256
metric space, 237
  compact sets, 256
  complete, 244
  completely separable, 244
  open set, 247
  separable, 246
migration matrix, 276
minimal polynomial
  finding it, 68
minimal polynomial
  algebraic number, 68
minimum polynomial, 68
  determined by measure, 68
  direct sum, 204
Minkowski inequality, 537
integrals, 537
Minkowski’s inequality, 540
minor, 125
mixed partial derivatives, 418
module
  finitely generated, 184
  mollifier, 545
  convolution, 545
monotone convergence theorem, 490
Moore-Penrose inverse, 343
  least squares, 344
  uniqueness, 344
morphism, 195
multi-index, 113
multi-index notation, 113
negative definite, 634
negative part, 635
Neumann series, 426
Neumann series, 426
Newton’s method, 420
nilpotent, 150
  block diagonal matrix, 152
  Jordan form, uniqueness, 152
  Jordan normal form, 152
Noetherian module, 188
Noetherian rings, 188
non equal mixed partials
  example, 419
  nondefective, 132
norm
  \(p\) norm, 256
  strictly convex, 369
  uniformly convex, 369
normal closure, 233
normal extension, 233
normal matrix, 312
null and rank, 306
open ball, 237
open set, 237
open cover, 237
open set, 247
open sets
  countable basis, 244
operator norm, 261
order, 15
orthonormal, 258
orthonormal polynomials, 303
outer measure, 476
  determined by measure, 476
measurable, 477
outer measure on \( \mathbb{R} \), 478
outer regular, 475
outer regularity, 481

p.i.d. relatively prime, 186
parallelepiped, 590
volume, 300, 590
partial derivatives, 408, 416
continuous, 417
partially ordered set, 603
partition of unity, 506
infinitely differentiable, 546
partitioned matrix, 89
Penrose conditions, 344
permutation, 166
permutation matrices, 92
Perron's theorem, 280
pivot, 43
pivot column, 39
pivot columns, 39
pivot position, 39
pivot positions, 39
points of density, 553
pointwise convergence, 249
polar decomposition, 570
polar form complex number, 10
Polish space, 244, 468
polynomial
addition, 16
degree, 16
divides, 17
division, 17
equality, 17
greatest common divisor, 17
greatest common divisor, uniqueness, 17
irreducible, 18
irreducible factorization, 18
multiplication, 18
relatively prime, 18
polynomial
leading term, 16
matrix coefficients, 182
monic, 18
polynomials
canceling, 19
factoring, 12
factorization, 19, 124
positive definite
positive eigenvalues, 338
principle minors, 338
positive definite matrix, 691
positive part, 493
positive definite, 338
power method, 377
prime number, 22
primitive, 442
principal ideal domain, 185

greatest common divisor, 186
principle minors, 338
projection map
convex set, 506
QR algorithm, 318, 387
convergence, 318
convergence theorem, 318
non convergence, 318
QR factorization, 318
quadratic form, 313
quadratic formula, 12
quotient
prime ideal, 194
quotient module, 194
quotient ring
field, 194
quotient space, 194
quotient vector space, 194
radical, 194
Radon measure, 244, 468
Radon Nikodym derivative, 571
Radon Nikodym Theorem
\( \sigma \) finite measures, 571
finite measures, 571
random variable
distribution measure, 470
rank
number of pivot columns, 88
rank of a matrix, 88, 174
rank theorem, 174
rational canonical form, 149, 205
Rayleigh quotient, 485
how close?, 485
real and imaginary parts, 493
regression line, 299
regular, 185
regular measure space, 321
regular Sturm Liouville problem, 304
relatively prime, 186
Riesz map, 562
Riesz representation theorem, 295
Hilbert space, 571
metric space, 506
Riesz Representation theorem
\( C (X) \), 580
Riesz representation theorem \( L^p \)
finite measures, 571
unitary, 297, 309
Unitary matrix
    representation, 297
upper Hessenberg matrix, 307
Vandermonde determinant, 182
variation of constants formula, 323
variational inequality, 305
vector measures, 51
vector space, 51
    axioms, 76
    dimension, 252
vector space axioms, 51
vector valued function
    limit theorems, 264
vectors, 53, 76
Vitali cover, 521
volume
    parallelepiped, 300
Weierstrass approximation
    estimate, 593
well ordered, 6
well ordered sets, 605
well ordering, 6
Wilson’s theorem, 30
Wronskian, 180, 322
Wronskian alternative, 322
Young’s inequality, 535, 583