# Contents

## I Basic Linear Algebra

1 Fundamentals .......................... 13
   1.0.1 Outcomes ............................. 13
   1.1 $\mathbb{R}^n$ ............................ 13
   1.2 Algebra in $\mathbb{R}^n$ ................. 15
   1.3 Lines .................................. 16
   1.4 Exercises ............................... 18

2 Systems Of Equations ................. 21
   2.1 Geometric Interpretations .............. 21
   2.2 Systems Of Equations, Algebraic Procedures ....................... 22
      2.2.1 Elementary Operations ................. 22
      2.2.2 Gauss Elimination .......................... 25
   2.3 Exercises ............................... 33

3 Matrices ................................ 39
   3.0.1 Outcomes ................................ 39
   3.1 Matrix Arithmetic ....................... 39
      3.1.1 Addition And Scalar Multiplication Of Matrices ................ 39
      3.1.2 Multiplication Of Matrices .................... 42
      3.1.3 The $ij$th Entry Of A Product .................... 45
      3.1.4 Properties Of Matrix Multiplication ............... 46
      3.1.5 The Transpose ............................ 47
      3.1.6 The Identity And Inverses ...................... 48
      3.1.7 Finding The Inverse Of A Matrix ................. 50
   3.2 Exercises ............................... 53

4 Determinants .......................... 57
   4.0.1 Outcomes ................................ 57
   4.1 Basic Techniques And Properties ............. 57
      4.1.1 Cofactors And $2 \times 2$ Determinants .................. 57
      4.1.2 The Determinant Of A Triangular Matrix .............. 60
      4.1.3 Properties Of Determinants ...................... 62
      4.1.4 Finding Determinants Using Row Operations ............. 63
   4.2 Applications ............................ 65
      4.2.1 A Formula For The Inverse ...................... 65
      4.2.2 Cramer’s Rule ............................. 68
   4.3 Exercises ............................... 70
   4.4 The Mathematical Theory Of Determinants ............. 75
   4.5 The Cayley Hamilton Theorem .................. 86
4.6 Exercises ................................................................. 88

5 Vector Spaces ......................................................... 91
  5.0.1 Outcomes ......................................................... 91
  5.1 Vector Spaces ..................................................... 91
    5.1.1 Vector Space Axioms ........................................ 91
    5.1.2 Spans ......................................................... 94
    5.1.3 Subspaces .................................................... 97
    5.1.4 Linear Independence ....................................... 98
    5.1.5 Basis And Dimension ................................. 100
    5.1.6 Proof Of Exchange Theorem .............................. 105
  5.2 Exercises ......................................................... 106

II Vector Calculus ................................................... 111

6 Vectors And Points In $\mathbb{R}^n$ ................................ 113
  6.0.1 Outcomes ....................................................... 113
  6.1 Distance in $\mathbb{R}^n$ ........................................ 113
  6.2 Open And Closed Sets ......................................... 117
  6.3 Exercises ......................................................... 120
  6.4 Physical Vectors ................................................. 122
  6.5 Exercises ......................................................... 126

7 Vector Products ..................................................... 129
  7.0.1 Outcomes ....................................................... 129
  7.1 The Dot Product ................................................ 129
  7.2 The Geometric Significance Of The Dot Product ....... 132
    7.2.1 The Angle Between Two Vectors ......................... 132
    7.2.2 Work And Projections .................................... 134
    7.2.3 The Parabolic Mirror, An Application .............. 136
    7.2.4 The Dot Product And Distance In $\mathbb{C}^n$ ....... 138
  7.3 Exercises ......................................................... 141
  7.4 The Cross Product .............................................. 142
    7.4.1 The Distributive Law For The Cross Product ........ 145
    7.4.2 Torque ....................................................... 146
    7.4.3 Center Of Mass ............................................ 147
    7.4.4 Angular Velocity ......................................... 148
    7.4.5 The Box Product .......................................... 150
  7.5 Vector Identities And Notation .............................. 151
  7.6 Exercises ......................................................... 154

8 Bases For $\mathbb{R}^n$ ............................................... 157
  8.0.1 Outcomes ....................................................... 157
  8.1 Orthonormal Bases .............................................. 157
    8.1.1 The Least Squares Regression Line .................... 160
    8.1.2 The Fredholm Alternative .............................. 161
  8.2 The Dual Basis ................................................. 162
  8.3 Exercises ......................................................... 167
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Linear Transformations</td>
<td>169</td>
</tr>
<tr>
<td>9.0.1</td>
<td>Outcomes</td>
<td>169</td>
</tr>
<tr>
<td>9.1</td>
<td>Linear Transformations</td>
<td>169</td>
</tr>
<tr>
<td>9.2</td>
<td>Constructing The Matrix Of A Linear Transformation</td>
<td>170</td>
</tr>
<tr>
<td>9.2.1</td>
<td>Rotations Of $\mathbb{R}^2$</td>
<td>171</td>
</tr>
<tr>
<td>9.2.2</td>
<td>Projections</td>
<td>172</td>
</tr>
<tr>
<td>9.2.3</td>
<td>Matrices Which Are One To One Or Onto</td>
<td>173</td>
</tr>
<tr>
<td>9.2.4</td>
<td>The General Solution Of A Linear System</td>
<td>175</td>
</tr>
<tr>
<td>9.3</td>
<td>Exercises</td>
<td>177</td>
</tr>
<tr>
<td>10</td>
<td>Spectral Theory</td>
<td>181</td>
</tr>
<tr>
<td>10.0.1</td>
<td>Outcomes</td>
<td>181</td>
</tr>
<tr>
<td>10.1</td>
<td>Eigenvalues And Eigenvectors Of A Matrix</td>
<td>181</td>
</tr>
<tr>
<td>10.1.1</td>
<td>Definition Of Eigenvectors And Eigenvalues</td>
<td>181</td>
</tr>
<tr>
<td>10.1.2</td>
<td>Finding Eigenvectors And Eigenvalues</td>
<td>183</td>
</tr>
<tr>
<td>10.1.3</td>
<td>A Warning</td>
<td>185</td>
</tr>
<tr>
<td>10.1.4</td>
<td>Complex Eigenvalues</td>
<td>187</td>
</tr>
<tr>
<td>10.2</td>
<td>Volumes</td>
<td>188</td>
</tr>
<tr>
<td>10.3</td>
<td>Block Multiplication Of Matrices</td>
<td>191</td>
</tr>
<tr>
<td>10.4</td>
<td>Shur’s Theorem</td>
<td>192</td>
</tr>
<tr>
<td>10.5</td>
<td>Exercises</td>
<td>197</td>
</tr>
<tr>
<td>11</td>
<td>Planes, And Surfaces In $\mathbb{R}^n$</td>
<td>201</td>
</tr>
<tr>
<td>11.0.1</td>
<td>Outcomes</td>
<td>201</td>
</tr>
<tr>
<td>11.1</td>
<td>Planes</td>
<td>201</td>
</tr>
<tr>
<td>11.2</td>
<td>Quadric Surfaces</td>
<td>204</td>
</tr>
<tr>
<td>11.3</td>
<td>Exercises</td>
<td>207</td>
</tr>
<tr>
<td>12</td>
<td>Vector Valued Functions</td>
<td>209</td>
</tr>
<tr>
<td>12.0.1</td>
<td>Outcomes</td>
<td>209</td>
</tr>
<tr>
<td>12.1</td>
<td>Vector Valued Functions</td>
<td>209</td>
</tr>
<tr>
<td>12.2</td>
<td>Vector Fields</td>
<td>210</td>
</tr>
<tr>
<td>12.3</td>
<td>Continuous Functions</td>
<td>212</td>
</tr>
<tr>
<td>12.3.1</td>
<td>Sufficient Conditions For Continuity</td>
<td>212</td>
</tr>
<tr>
<td>12.4</td>
<td>Limits Of A Function</td>
<td>213</td>
</tr>
<tr>
<td>12.5</td>
<td>Properties Of Continuous Functions</td>
<td>216</td>
</tr>
<tr>
<td>12.6</td>
<td>Exercises</td>
<td>217</td>
</tr>
<tr>
<td>12.7</td>
<td>Some Fundamentals</td>
<td>219</td>
</tr>
<tr>
<td>12.7.1</td>
<td>The Nested Interval Lemma</td>
<td>222</td>
</tr>
<tr>
<td>12.7.2</td>
<td>The Extreme Value Theorem</td>
<td>223</td>
</tr>
<tr>
<td>12.7.3</td>
<td>Sequences And Completeness</td>
<td>224</td>
</tr>
<tr>
<td>12.7.4</td>
<td>Continuity And The Limit Of A Sequence</td>
<td>227</td>
</tr>
<tr>
<td>12.8</td>
<td>Exercises</td>
<td>227</td>
</tr>
<tr>
<td>13</td>
<td>Vector Valued Functions Of One Variable</td>
<td>231</td>
</tr>
<tr>
<td>13.0.1</td>
<td>Outcomes</td>
<td>231</td>
</tr>
<tr>
<td>13.1</td>
<td>Limits Of A Vector Valued Function Of One Variable</td>
<td>231</td>
</tr>
<tr>
<td>13.2</td>
<td>The Derivative And Integral</td>
<td>233</td>
</tr>
<tr>
<td>13.2.1</td>
<td>Geometric And Physical Significance Of The Derivative</td>
<td>234</td>
</tr>
<tr>
<td>13.2.2</td>
<td>Differentiation Rules</td>
<td>236</td>
</tr>
<tr>
<td>13.2.3</td>
<td>Leibniz’s Notation</td>
<td>238</td>
</tr>
<tr>
<td>13.3</td>
<td>Product Rule For Matrices</td>
<td>238</td>
</tr>
</tbody>
</table>
CONTENTS

13.4 Moving Coordinate Systems ........................................... 239
13.5 Exercises ............................................................... 241
13.6 Newton’s Laws Of Motion .............................................. 243
  13.6.1 Kinetic Energy ................................................... 247
  13.6.2 Impulse And Momentum ......................................... 247
13.7 Moving Coordinate Systems .......................................... 248
  13.7.1 The Coriolis Acceleration ..................................... 248
  13.7.2 The Coriolis Acceleration On The Rotating Earth .......... 250
13.8 Exercises ............................................................... 255
13.9 Line Integrals ......................................................... 258
  13.9.1 Arc Length And Orientations ................................ 258
  13.9.2 Line Integrals And Work ...................................... 260
  13.9.3 Another Notation For Line Integrals ......................... 263
13.10 Exercises .............................................................. 264

14 Motion On A Space Curve ............................................. 265
  14.0.1 Outcomes .......................................................... 265
14.1 Space Curves .......................................................... 265
14.2 Exercises ............................................................... 269
14.3 Independence Of Parameterization∗ ............................... 270
  14.3.1 Hard Calculus ................................................... 270
  14.3.2 Independence Of Parameterization ......................... 274

15 Some Curvilinear Coordinate Systems ............................. 277
  15.0.3 Outcomes .......................................................... 277
15.1 Polar Cylindrical And Spherical Coordinates .................... 277
15.2 The Acceleration In Polar Coordinates ......................... 279
15.3 Planetary Motion ..................................................... 282
15.4 Exercises ............................................................... 286

16 Functions Of Many Variables ....................................... 289
  16.0.1 Outcomes .......................................................... 289
16.1 The Graph Of A Function Of Two Variables ...................... 289
16.2 Review Of Limits ..................................................... 291
16.3 The Directional Derivative And Partial Derivatives .......... 292
  16.3.1 The Directional Derivative .................................. 292
  16.3.2 Partial Derivatives .......................................... 294
16.4 Mixed Partial Derivatives ......................................... 296
16.5 Partial Differential Equations ................................... 298
16.6 Exercises ............................................................... 298

17 The Derivative Of A Function Of Many Variables ............. 301
  17.0.1 Outcomes .......................................................... 301
17.1 The Derivative Of Functions Of One Variable ................. 301
17.2 The Derivative Of Functions Of Many Variables .............. 303
17.3 C1 Functions .......................................................... 304
  17.3.1 Approximation With A Tangent Plane ..................... 309
17.4 The Chain Rule .......................................................... 310
17.5 Lagrangian Mechanics∗ ............................................ 316
17.6 Newton’s Method ...................................................... 321
  17.6.1 The Newton Raphson Method In One Dimension .......... 321
  17.6.2 Newton’s Method For Nonlinear Systems ................. 322
## CONTENTS

17.7 Convergence Questions*  
17.7.1 A Fixed Point Theorem  
17.7.2 The Operator Norm  
17.7.3 A Method For Finding Zeros  
17.7.4 Newton’s Method  
17.8 Exercises  

18 The Gradient  
18.0.1 Outcomes  
18.1 Fundamental Properties  
18.2 Tangent Planes  
18.3 Exercises  

19 Optimization  
19.0.1 Outcomes  
19.1 Local Extrema  
19.2 The Second Derivative Test  
19.3 Proof Of The Second Derivative Test  
19.4 Exercises  
19.5 Lagrange Multipliers  
19.6 Exercises  

20 The Riemann Integral On $\mathbb{R}^n$  
20.0.1 Outcomes  
20.1 Methods For Double Integrals  
20.1.1 Density And Mass  
20.2 Exercises  
20.3 Methods For Triple Integrals  
20.3.1 Definition Of The Integral  
20.3.2 Iterated Integrals  
20.3.3 Mass And Density  
20.4 Exercises With Answers  
20.5 Exercises  

21 The Integral In Other Coordinates  
21.0.1 Outcomes  
21.1 Different Coordinates  
21.2 Exercises With Answers  
21.3 Exercises  
21.4 The Moment Of Inertia  
21.4.1 The Spinning Top  
21.4.2 Kinetic Energy  
21.4.3 Finding The Moment Of Inertia And Center Of Mass  
21.5 Exercises  

22 The Integral On Other Sets  
22.0.1 Outcomes  
22.1 The $p$ Dimensional Volume In $\mathbb{R}^n$  
22.2 Spherical Coordinates In $\mathbb{R}^n$  
22.3 Exercises With Answers  
22.4 Exercises
23 Calculus Of Vector Fields

23.0.1 Outcomes ........................................... 421
23.1 Divergence And Curl Of A Vector Field .................. 421
23.1.1 Vector Identities .................................. 422
23.1.2 Vector Potentials ................................. 424
23.1.3 The Weak Maximum Principle ....................... 424
23.2 Exercises ............................................. 425
23.3 The Divergence Theorem ............................... 426
23.3.1 Coordinate Free Concept Of Divergence ............ 429
23.4 Some Applications Of The Divergence Theorem ........... 430
23.4.1 Hydrostatic Pressure ............................. 430
23.4.2 Archimedes Law Of Buoyancy .................... 431
23.4.3 Equations Of Heat And Diffusion .................. 431
23.4.4 Balance Of Mass ................................ 432
23.4.5 Balance Of Momentum ............................. 433
23.4.6 The Wave Equation ............................... 438
23.4.7 A Negative Observation ............................ 438
23.4.8 Volumes Of Balls In $\mathbb{R}^n$ (For Those Who Know About The Gamma Function) .................... 438
23.4.9 Electrostatics ...................................... 441
23.5 Exercises ............................................. 442

24 Stokes And Green's Theorems ............................... 445

24.0.1 Outcomes ........................................... 445
24.1 Green’s Theorem ..................................... 445
24.2 Stoke’s Theorem From Green’s Theorem .................. 449
24.2.1 Orientation ....................................... 452
24.2.2 Conservative Vector Fields ....................... 453
24.2.3 Some Terminology ................................ 456
24.2.4 Maxwell’s Equations And The Wave Equation ........ 456
24.3 Exercises ............................................. 458

25 Curvilinear Coordinates ................................... 461

25.0.1 Outcomes ........................................... 461
25.1 Exercises ............................................. 464
25.2 Transformation Of Coordinates .......................... 467
25.3 Differentiation And Christoffel Symbols ............... 468
25.4 Gradients And Divergence ............................. 470
25.5 Exercises ............................................. 472
25.6 Curl And Cross Products ............................... 473

26 The Theory Of The Riemann Integral* ....................... 477

26.1 Basic Properties ...................................... 479
26.2 Iterated Integrals .................................... 492
26.2.1 Some Observations ............................... 496

III Further Topics In Linear Algebra ............................ 497

27 Rank Of A Matrix ......................................... 499

27.0.2 Outcomes ........................................... 499
27.1 The Row Reduced Echelon Form Of A Matrix .......... 499
CONTENTS

27.2 The Rank Of A Matrix .................................................. 503
  27.2.1 The Definition Of Rank ............................................. 503
  27.2.2 Finding The Rank Of A Matrix .................................... 503
  27.2.3 Rank And Existence Of Solutions To Linear Systems .......... 506
27.3 Exercises ................................................................. 506

28 The \textit{LU} Decomposition ............................................... 511
  28.0.1 Outcomes ............................................................. 511
  28.1 Definition Of An \textit{LU} Decomposition ......................... 511
  28.2 Finding An \textit{LU} Decomposition .................................. 511
  28.3 Solving Systems Using The \textit{LU} Decomposition ............... 513
  28.4 The \textit{PLU} Decomposition ......................................... 514
  28.5 Exercises ............................................................... 516

29 Applications Of Spectral Theory ....................................... 519
  29.0.1 Outcomes ............................................................. 519
  29.1 Defective And Nondefective Matrices ............................. 519
  29.2 Some Applications Of Eigenvalues And Eigenvectors ........... 522
    29.2.1 Principle Directions ............................................. 522
    29.2.2 Migration Matrices .............................................. 524
  29.3 The Estimation Of Eigenvalues ..................................... 527
  29.4 Exercises ............................................................... 528

30 Some Special Matrices .................................................. 533
  30.0.1 Outcomes ............................................................. 533
  30.1 Symmetric And Orthogonal Matrices ................................ 533
    30.1.1 Orthogonal Matrices ............................................. 533
    30.1.2 Symmetric And Skew Symmetric Matrices .................... 535
    30.1.3 Diagonalizing A Symmetric Matrix ............................ 542
  30.2 More General Considerations ....................................... 543
  30.3 Exercises ............................................................... 543

31 Numerical Methods For Solving Linear Systems ...................... 549
  31.0.1 Outcomes ............................................................. 549
  31.1 Iterative Methods For Linear Systems ............................ 549
    31.1.1 The Jacobi Method .............................................. 550
    31.1.2 The Gauss Seidel Method ....................................... 552
  31.2 Exercises ............................................................... 556

32 Numerical Methods For Solving The Eigenvalue Problem ............ 559
  32.0.1 Outcomes ............................................................. 559
  32.1 The Power Method For Eigenvalues ................................ 559
  32.2 The Shifted Inverse Power Method ................................ 562
    32.2.1 Complex Eigenvalues ........................................... 571
  32.3 The Rayleigh Quotient ............................................... 573
  32.4 Exercises ............................................................... 576
## CONTENTS

### A Worked Exercises For Linear Algebra
- A.1 Worked Exercises ................................................. 579
- A.2 Worked Exercises ................................................. 581
- A.3 Worked Exercises ................................................. 583
- A.4 Worked Exercises ................................................. 586
- A.5 Worked Exercises ................................................. 590
- A.6 Worked Exercises ................................................. 594
- A.7 Worked Exercises ................................................. 596
- A.8 Worked Exercises ................................................. 597

### B The Fundamental Theorem Of Algebra .......................... 601

### C Heroic Linear Algebra ............................................. 603
- C.1 Vector Spaces .................................................... 603
- C.2 Matrix Multiplication As A Linear Transformation ........ 608
- C.3 \( \mathcal{L}(V, W) \) As A Vector Space ......................... 608
- C.4 Eigenvalues And Eigenvectors Of Linear Transformations ... 609
- C.5 Block Diagonal Matrices ......................................... 614
- C.6 The Matrix Of A Linear Transformation ....................... 618
- C.7 The Jordan Canonical Form ..................................... 624
- C.8 Convergence ....................................................... 630
  - C.8.1 The Concept Of A Norm .................................... 630
  - C.8.2 The Operator Norm ........................................ 633
- C.9 The Spectral Radius ............................................... 637
- C.10 Convergence For Iterative Methods ......................... 640
Part I

Basic Linear Algebra
1.0.1 Outcomes

1. Describe $\mathbb{R}^n$ and do algebra with vectors in $\mathbb{R}^n$.
2. Represent a line in 3 space by a vector parameterization, a set of scalar parametric equations or using symmetric form.
3. Find a parameterization of a line given information about
   
   (a) a point of the line and the direction of the line
   (b) two points contained in the line
4. Determine the direction of a line given its parameterization.

1.1 $\mathbb{R}^n$

The notation, $\mathbb{R}^n$, refers to the collection of ordered lists of $n$ real numbers. More precisely, consider the following definition.

**Definition 1.1.1** Define

$$\mathbb{R}^n \equiv \{(x_1, \cdots, x_n) : x_j \in \mathbb{R} \text{ for } j = 1, \cdots, n\}.$$ 

$(x_1, \cdots, x_n) = (y_1, \cdots, y_n)$ if and only if for all $j = 1, \cdots, n$, $x_j = y_j$. When $(x_1, \cdots, x_n) \in \mathbb{R}^n$, it is conventional to denote $(x_1, \cdots, x_n)$ by the single bold face letter, $x$. The numbers, $x_j$ are called the coordinates. The set

$$\{(0, \cdots, 0, t, 0, \cdots, 0) : t \in \mathbb{R} \}$$

for $t$ in the $i^{th}$ slot is called the $i^{th}$ coordinate axis. The point $0 \equiv (0, \cdots, 0)$ is called the origin.

Thus $(1, 2, 4) \in \mathbb{R}^3$ and $(2, 1, 4) \in \mathbb{R}^3$ but $(1, 2, 4) \neq (2, 1, 4)$ because, even though the same numbers are involved, they don’t match up. In particular, the first entries are not equal.

Why would anyone be interested in such a thing? First consider the case when $n = 1$. Then from the definition, $\mathbb{R}^1 = \mathbb{R}$. Recall that $\mathbb{R}$ is identified with the points of a line. Look at the number line again. Observe that this amounts to identifying a point on this line with a real number. In other words a real number determines where you are on this line. Now suppose $n = 2$ and consider two lines which intersect each other at right angles as shown in the following picture.
Notice how you can identify a point shown in the plane with the ordered pair, \((2, 6)\). You go to the right a distance of 2 and then up a distance of 6. Similarly, you can identify another point in the plane with the ordered pair \((-8, 3)\). Go to the left a distance of 8 and then up a distance of 3. The reason you go to the left is that there is a \(-\) sign on the eight. From this reasoning, every ordered pair determines a unique point in the plane. Conversely, taking a point in the plane, you could draw two lines through the point, one vertical and the other horizontal and determine unique points, \(x_1\) on the horizontal line in the above picture and \(x_2\) on the vertical line in the above picture, such that the point of interest is identified with the ordered pair, \((x_1, x_2)\). In short, points in the plane can be identified with ordered pairs similar to the way that points on the real line are identified with real numbers. Now suppose \(n = 3\). As just explained, the first two coordinates determine a point in a plane. Letting the third component determine how far up or down you go, depending on whether this number is positive or negative, this determines a point in space. Thus, \((1, 4, -5)\) would mean to determine the point in the plane that goes with \((1, 4)\) and then to go below this plane a distance of 5 to obtain a unique point in space. You see that the ordered triples correspond to points in space just as the ordered pairs correspond to points in a plane and single real numbers correspond to points on a line.

You can’t stop here and say that you are only interested in \(n \leq 3\). What if you were interested in the motion of two objects? You would need three coordinates to describe where the first object is and you would need another three coordinates to describe where the other object is located. Therefore, you would need to be considering \(\mathbb{R}^6\). If the two objects moved around, you would need a time coordinate as well. As another example, consider a hot object which is cooling and suppose you want the temperature of this object. How many coordinates would be needed? You would need one for the temperature, three for the position of the point in the object and one more for the time. Thus you would need to be considering \(\mathbb{R}^5\). Many other examples can be given. Sometimes \(n\) is very large. This is often the case in applications to business when they are trying to maximize profit subject to constraints. It also occurs in numerical analysis when people try to solve hard problems on a computer.

There are other ways to identify points in space with three numbers but the one presented is the most basic. In this case, the coordinates are known as Cartesian coordinates after Descartes\(^1\) who invented this idea in the first half of the seventeenth century. I will often not bother to draw a distinction between the point in \(n\) dimensional space and its Cartesian coordinates.

\(^1\)René Descartes 1596-1650 is often credited with inventing analytic geometry although it seems the ideas were actually known much earlier. He was interested in many different subjects, physiology, chemistry, and physics being some of them. He also wrote a large book in which he tried to explain the book of Genesis scientifically. Descartes ended up dying in Sweden.
There are two algebraic operations done with elements of $\mathbb{R}^n$. One is addition and the other is multiplication by numbers, called scalars.

**Definition 1.2.1** If $x \in \mathbb{R}^n$ and $a$ is a number, also called a scalar. Then $ax \in \mathbb{R}^n$ is defined by

$$ a \cdot x = a(x_1, \ldots, x_n) \equiv (ax_1, \ldots, ax_n). \quad (1.1) $$

This is known as scalar multiplication. If $x, y \in \mathbb{R}^n$ then $x + y \in \mathbb{R}^n$ and is defined by

$$ x + y = (x_1, \ldots, x_n) + (y_1, \ldots, y_n) \equiv (x_1 + y_1, \ldots, x_n + y_n) \quad (1.2) $$

With this definition, the algebraic properties satisfy the conclusions of the following theorem.

**Theorem 1.2.2** For $v, w \in \mathbb{R}^n$ and $\alpha, \beta$ scalars, (real numbers), the following hold.

$$ v + w = w + v, \quad (1.3) $$

the commutative law of addition,

$$ (v + w) + z = v + (w + z), \quad (1.4) $$

the associative law for addition,

$$ v + 0 = v, \quad (1.5) $$

the existence of an additive identity,

$$ v + (-v) = 0, \quad (1.6) $$

the existence of an additive inverse, Also

$$ \alpha (v + w) = \alpha v + \alpha w, \quad (1.7) $$

$$ (\alpha + \beta) v = \alpha v + \beta v, \quad (1.8) $$

$$ \alpha (\beta v) = \alpha \beta (v), \quad (1.9) $$

$$ 1v = v. \quad (1.10) $$

In the above $0 = (0, \ldots, 0)$.

You should verify these properties all hold. For example, consider (1.7)

$$ \alpha (v + w) = \alpha (v_1 + w_1, \ldots, v_n + w_n) $$
$$ = (\alpha v_1 + \alpha w_1, \ldots, \alpha v_n + \alpha w_n) $$
$$ = (\alpha v_1, \ldots, \alpha v_n) + (\alpha w_1, \ldots, \alpha w_n) $$
$$ = \alpha v + \alpha w. $$

As usual subtraction is defined as $x - y \equiv x + (-y)$. 
1.3 Lines

To begin with consider the case $n = 1, 2$. In the case where $n = 1$, the only line is just $\mathbb{R}^1 = \mathbb{R}$. Therefore, if $x_1$ and $x_2$ are two different points in $\mathbb{R}$, consider

$$x = x_1 + t (x_2 - x_1)$$

where $t \in \mathbb{R}$ and the totality of all such points will give $\mathbb{R}$. You see that you can always solve the above equation for $t$, showing that every point on $\mathbb{R}$ is of this form. Now consider the plane. Does a similar formula hold? Let $(x_1, y_1)$ and $(x_2, y_2)$ be two different points in $\mathbb{R}^2$ which are contained in a line, $l$. Suppose that $x_1 \neq x_2$. Then if $(x, y)$ is an arbitrary point on $l$,

Now by similar triangles,

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{y - y_1}{x - x_1}$$

and so the point slope form of the line, $l$, is given as

$$y - y_1 = m (x - x_1).$$

If $t$ is defined by

$$x = x_1 + t (x_2 - x_1),$$

you obtain this equation along with

$$y = y_1 + mt (x_2 - x_1) = y_1 + t (y_2 - y_1).$$

Therefore,

$$(x, y) = (x_1, y_1) + t (x_2 - x_1, y_2 - y_1).$$

If $x_1 = x_2$, then in place of the point slope form above, $x = x_1$. Since the two given points are different, $y_1 \neq y_2$ and so you still obtain the above formula for the line. Because of this, the following is the definition of a line in $\mathbb{R}^n$.

**Definition 1.3.1** A line in $\mathbb{R}^n$ containing the two different points, $x^1$ and $x^2$ is the collection of points of the form

$$x = x^1 + t (x^2 - x^1)$$

where $t \in \mathbb{R}$. This is known as a **parametric equation** and the variable $t$ is called the **parameter**.
Often $t$ denotes time in applications to Physics. Note this definition agrees with the usual notion of a line in two dimensions and so this is consistent with earlier concepts.

**Lemma 1.3.2** Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ with $\mathbf{a} \neq \mathbf{0}$. Then $\mathbf{x} = t\mathbf{a} + \mathbf{b}, t \in \mathbb{R}$, is a line.

**Proof**: Let $\mathbf{x}^1 = \mathbf{b}$ and let $\mathbf{x}^2 - \mathbf{x}^1 = \mathbf{a}$ so that $\mathbf{x}^2 \neq \mathbf{x}^1$. Then $t\mathbf{a} + \mathbf{b} = \mathbf{x}^1 + t(\mathbf{x}^2 - \mathbf{x}^1)$ and so $\mathbf{x} = t\mathbf{a} + \mathbf{b}$ is a line containing the two different points, $\mathbf{x}^1$ and $\mathbf{x}^2$. This proves the lemma.

**Definition 1.3.3** The vector $\mathbf{a}$ in the above lemma is called a direction vector for the line.

**Definition 1.3.4** Let $\mathbf{p}$ and $\mathbf{q}$ be two points in $\mathbb{R}^n$, $\mathbf{p} \neq \mathbf{q}$. The directed line segment from $\mathbf{p}$ to $\mathbf{q}$, denoted by $\overrightarrow{\mathbf{p}\mathbf{q}}$, is defined to be the collection of points,

$$\mathbf{x} = \mathbf{p} + t(\mathbf{q} - \mathbf{p}), \ t \in [0, 1]$$

with the direction corresponding to increasing $t$.

Think of $\overrightarrow{\mathbf{p}\mathbf{q}}$ as an arrow whose point is on $\mathbf{q}$ and whose base is at $\mathbf{p}$ as shown in the following picture.

![Diagram of directed line segment](image)

This line segment is a part of a line from the above Definition.

**Example 1.3.5** Find a parametric equation for the line through the points $(1, 2, 0)$ and $(2, -4, 6)$.

Use the definition of a line given above to write

$$(x, y, z) = (1, 2, 0) + t(1, -6, 6), \ t \in \mathbb{R}.$$  

The reason for the word, “a”, rather than the word, “the” is there are infinitely many different parametric equations for the same line. To see this replace $t$ with $3s$. Then you obtain a parametric equation for the same line because the same set of points are obtained. The difference is they are obtained from different values of the parameter. What happens is this: The line is a set of points but the parametric description gives more information than that. It tells us how the set of points are obtained. Obviously, there are many ways to trace out a given set of points and each of these ways corresponds to a different parametric equation for the line.

**Example 1.3.6** Find a parametric equation for the line which contains the point $(1, 2, 0)$ and has direction vector, $(1, 2, 1)$.

From the above this is just

$$(x, y, z) = (1, 2, 0) + t(1, 2, 1), \ t \in \mathbb{R}.$$  

(1.11)
Sometimes people elect to write a line like the above in the form
\[ x = 1 + t, \ y = 2 + 2t, \ z = t, \ t \in \mathbb{R}. \] (1.12)

This is a set of scalar parametric equations which amounts to the same thing as (1.11).

There is one other form for a line which is sometimes considered useful. It is the so-called symmetric form. Consider the line of (1.12). You can solve for the parameter, \( t \) to write
\[ t = x - 1, \ t = \frac{y - 2}{2}, \ t = z. \]

Therefore,
\[ x - 1 = \frac{y - 2}{2} = z. \]

This is the symmetric form of the line.

**Example 1.3.7** Suppose the symmetric form of a line is
\[ \frac{x - 2}{3} = \frac{y - 1}{2} = z + 3. \]

Find the line in parametric form.

Let \( t = \frac{x - 2}{3}, \ t = \frac{y - 1}{2} \) and \( t = z + 3 \). Then solving for \( x, y, z \), you get
\[ x = 3t + 2, \ y = 2t + 1, \ z = t - 3, \ t \in \mathbb{R}. \]

Written in terms of vectors this is
\[ (2, 1, -3) + t (3, 2, 1) = (x, y, z), \ t \in \mathbb{R}. \]

### 1.4 Exercises

1. Verify all the properties (1.3)-(1.10).
2. Compute the following
   (a) \( 5 (1, 2, 3, -2) + 6 (2, 1, -2, 7) \)
   (b) \( 5 (1, 2, -2) - 6 (2, 1, -2) \)
   (c) \( -3 (1, 0, 3, -2) + (2, 0, -2, 1) \)
   (d) \( -3 (1, -2, -3, -2) - 2 (2, -1, -2, 7) \)
   (e) \( - (2, 2, -3, -2) + 2 (2, 4, -2, 7) \)
3. Find a parametric equation for the line through the points \((2, 3, 4)\) and \((-2, 3, 0)\).
4. Find a parametric equation for the line through the points \((2, 0, 4)\) and \((-2, 5, 0)\).
5. Find symmetric equations for the line through the points \((2, 2, 4)\) and \((-2, 3, 1)\).
6. Find symmetric equations for the line through the points \((1, 2, 4)\) and \((-2, 1, 1)\).
7. Symmetric equations for a line are
\[ \frac{x + 1}{3} = \frac{2x + 3}{2} = z + 7. \]

Find parametric equations for this line.
8. Symmetric equations for a line are
\[
\frac{x - 1}{3} = \frac{2x - 3}{5} = z + 5.
\]
Find parametric equations for this line.

9. Parametric equations for a line are
\[
x = 1 + 2t, \quad y = 3 - t, \quad z = 5 + 3t.
\]
Find symmetric equations for this line. Parametric equations for a line are
\[
x = 1 + 4t, \quad y = 3 + t, \quad z = 5 - 3t.
\]
Find symmetric equations for this line.

10. Find the equation of the line through (1, 2, 3) having direction vector, (-3, 2, -4).

11. Find the equation of the line through (1, 0, 3) having direction vector, (1, 2, -4).

12. Parametric equations for a line are
\[
x = 1 + 2t, \quad y = 3 - t, \quad z = 5 + 3t.
\]
What is a direction vector for this line?

13. Parametric equations for a line are
\[
x = 1 + 8t, \quad y = 3 + t, \quad z = 7 + 2t.
\]
What is a direction vector for this line?
Systems Of Equations

1. Relate the types of solution sets of a system of two or three variables to the intersections of lines in a plane or the intersection of planes in three space.

2. Determine whether a system of linear equations has no solution, a unique solution or an infinite number of solutions from its echelon form.

3. Solve a system of equations using Gauss elimination.

4. Model a physical system with linear equations and then solve

2.1 Geometric Interpretations

As you know from high school, equations like $2x + 3y = 6$ can be graphed as straight lines. To find the solution to two such equations, you could graph the two straight lines and the ordered pairs identifying the point (or points) of intersection would give the $x$ and $y$ values of the solution to the two equations because such an ordered pair satisfies both equations. The following picture illustrates what can occur with two equations involving two variables.

In the first example of the above picture, there is a unique point of intersection. In the second, there are no points of intersection. The other thing which can occur is that the two lines are really the same line. For example, $x + y = 1$ and $2x + 2y = 2$ are relations which when graphed yield the same line. In this case there are infinitely many points in the simultaneous solution of these two equations, every ordered pair which is on the graph of the line. It is always this way when considering linear systems of equations. There is either no solution, exactly one or infinitely many although the reasons for this are not completely comprehended by considering a simple picture in two dimensions.

Example 2.1.1 Find the solution to the system $x + y = 3$, $y - x = 5$.

You can verify the solution is $(x, y) = (-1, 4)$. You can see this geometrically by graphing the equations of the two lines. If you do so correctly, you should obtain a graph which looks
something like the following in which the point of intersection represents the solution of the two equations.

\[(x, y) = (-1, 4)\]

Example 2.1.2 You can also imagine other situations such as the case of three intersecting lines having no common point of intersection or three intersecting lines which do intersect at a single point as illustrated in the following picture.

In the case of the first picture above, there would be no solution to the three equations whose graphs are the given lines. In the case of the second picture there is a solution to the three equations whose graphs are the given lines.

An equation like \(2x + 4y - 5z = 8\) involving three variables is a plane in three dimensions. This will be discussed later. Therefore, the geometrical significance of solving systems of equations involving three variables, is to take the intersection of planes.

Example 2.1.3 In the case of the intersection of planes, you can imagine that the intersection of two planes is a line and then if you have another plane, it could intersect this line in a single point or it could contain the line or be parallel to the line and not have any intersection with it.

In higher dimensions it is customary to refer to such relations like \(x + y - 2z + 4w = 8\) as a hyper-plane. Such pictures as above are useful in two or three dimensions for gaining insight into what can happen but they are not adequate for obtaining the exact solution set of the linear system. Furthermore, it is impossible to consider all possibilities through an attempt to draw pictures even in two or three dimensions and in higher dimensions, pictures are even less useful. The only rational and useful way to deal with this subject is through the use of algebra.

2.2 Systems Of Equations, Algebraic Procedures

2.2.1 Elementary Operations

Consider the following example.

Example 2.2.1 Find \(x\) and \(y\) such that

\[x + y = 7 \text{ and } 2x - y = 8.\]
The set of ordered pairs, \((x, y)\) which solve both equations is called the \textbf{solution set}.

You can verify that \((x, y) = (5, 2)\) 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 2.2.2** \textbf{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 2.2.3** 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 2.2.2 do not change the set of solutions to the system of linear equations.

**Theorem 2.2.4** Suppose you have two equations, involving the variables, \((x_1, \ldots, x_n)\)

\[E_1 = f_1, \ E_2 = f_2\]  \hspace{1cm} (2.2)

where \(E_1\) and \(E_2\) are expressions involving the variables. (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

\[E_1 = f_1, \ E_2 + aE_1 = f_2 + af_1.\]  \hspace{1cm} (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 \((x_1, \ldots, x_n)\) 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 \((x_1, \ldots, x_n)\) 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 $E_1 = f_1, E_2 = f_2$ has the same solution as the system, $E_2 = f_2, E_1 = f_1$ 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 2.2.5** Find the solutions to the system,

\[
\begin{align*}
x + 3y + 6z &= 25 \\
2x + 7y + 14z &= 58 \\
2y + 5z &= 19
\end{align*}
\]

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

\[
\begin{align*}
x + 3y + 6z &= 25 \\
y + 2z &= 8 \\
2y + 5z &= 19
\end{align*}
\]

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

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

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 (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.
2.2.2 Gauss Elimination

A less cumbersome way to represent a linear system is to write it as an augmented matrix. For example the linear system, (2.4) 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 \(x\) column, \[
\begin{pmatrix}
1 \\
2 \\
0
\end{pmatrix},
\]
a \(y\) column, \[
\begin{pmatrix}
3 \\
7 \\
2
\end{pmatrix},
\]
and a \(z\) column, \[
\begin{pmatrix}
6 \\
14 \\
5
\end{pmatrix}.
\]
The rows correspond to the 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 (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 (2.5). Next take \((-2)\) times the second row and add to the third,

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

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 (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

\[
a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\
\vdots \\
a_{m1}x_1 + \cdots + a_{mn}x_n = b_m
\]

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

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

Changes to the system of equations in (2.7) as a result of an elementary operations translate into changes of the augmented matrix resulting from a row operation. Note that Theorem 2.2.4 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 2.2.6 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 2.2.7 An augmented matrix is in 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.

Definition 2.2.8 An augmented matrix is in echelon form if all the above hold except that the entries above a leading entry are not required to equal zero.

Example 2.2.9 Here are some augmented matrices which are in reduced echelon form.

\[
\begin{pmatrix}
1 & 0 & 6 & 5 & 8 & | & 2 \\
0 & 1 & 2 & 7 & | & 3 \\
0 & 0 & 0 & 0 & | & 1 \\
0 & 0 & 0 & 0 & | & 0 \\
\end{pmatrix}, \quad \begin{pmatrix}
1 & 0 & 0 & | & 0 \\
0 & 2 & 0 & | & 0 \\
0 & 0 & 3 & | & 0 \\
0 & 0 & 0 & | & 1 \\
0 & 0 & 0 & | & 0 \\
\end{pmatrix}
\]

Example 2.2.10 Here are augmented matrices in echelon form which are not in reduced echelon form.

\[
\begin{pmatrix}
1 & 0 & 6 & 5 & 8 & | & 2 \\
0 & 1 & 2 & 7 & | & 3 \\
0 & 0 & 0 & 0 & | & 1 \\
0 & 0 & 0 & 0 & | & 0 \\
\end{pmatrix}, \quad \begin{pmatrix}
1 & 3 & 5 & | & 4 \\
0 & 2 & 0 & | & 7 \\
0 & 0 & 3 & | & 0 \\
0 & 0 & 0 & | & 1 \\
0 & 0 & 0 & | & 0 \\
\end{pmatrix}
\]

Example 2.2.11 Here are some augmented matrices which are not in echelon form.

\[
\begin{pmatrix}
0 & 0 & 0 & | & 0 \\
1 & 2 & 3 & | & 3 \\
0 & 2 & 0 & | & 2 \\
0 & 0 & 0 & | & 1 \\
0 & 0 & 0 & | & 0 \\
\end{pmatrix}, \quad \begin{pmatrix}
1 & 2 & & | & 3 \\
2 & 4 & & | & -6 \\
4 & 0 & & | & 7 \\
\end{pmatrix}, \quad \begin{pmatrix}
0 & 2 & 3 & | & 3 \\
1 & 5 & 0 & | & 2 \\
7 & 5 & 0 & | & 1 \\
0 & 0 & 1 & | & 0 \\
\end{pmatrix}
\]

Definition 2.2.12 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 2.2.13 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:

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

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

The row reduction algorithm

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 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. Repeat steps 1 and 2 for the matrix obtained by ignoring the row containing the pivot and the rows above along with the pivot column and the columns to the left of the pivot column.

4. Continue till a matrix in echelon form has been obtained.

5. Finally, 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.
When applying the algorithm, it is best to not make explicit mention of the lines dividing the last column from the rest of the matrix.

**Example 2.2.14** Here is a matrix:

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

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

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

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
\[
\begin{pmatrix}
0 & 1 & 1 & 4 & 0 \\
0 & 0 & 6 & 9 & 0 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
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.
\[
\begin{pmatrix}
0 & 3 & 3 & 0 & 0 \\
0 & 0 & 6 & 0 & 0 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
Next work on the fourth column in the same way.
\[
\begin{pmatrix}
0 & 3 & 0 & 0 & 0 \\
0 & 0 & 6 & 0 & 0 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
Finally, take \(-1/2\) times the second row and add to the first.
\[
\begin{pmatrix}
0 & 3 & 0 & 0 & 0 \\
0 & 0 & 6 & 0 & 0 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
This is now in reduced echelon form. You can put in the dividing lines between the matrix and the last column if you desire.
\[
\begin{pmatrix}
0 & 3 & 0 & 0 & \mid & 0 \\
0 & 0 & 6 & 0 & \mid & 0 \\
0 & 0 & 0 & -3 & \mid & 0 \\
0 & 0 & 0 & 0 & \mid & -3 \\
0 & 0 & 0 & 0 & \mid & 0
\end{pmatrix}
\]
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 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.

Example 2.2.15  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 2.2.16  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. 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 \textit{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 \(y = 10z\) and \(x = 3 + 5z\). Apparently \(z\) can equal any number. Lets 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 \textbf{basic variables}. The other variables are called \textbf{free variables}. In Example 2.2.16 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 2.2.16 this parameter was \(t\). Sometimes there are many free variables and in these cases, you need to use many parameters. Here is another example.

\textbf{Example 2.2.17} \textit{Find the solution to the system}

\[
x + 2y - z + w = 3 \\
x + y - z + w = 1 \\
x + 3y - z + w = 5
\]

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

\footnote{In this context \(t\) is called a \textit{parameter}.}
Now add the second row to the bottom row

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

\[
\begin{pmatrix}
x \\
y \\
z \\
w \\
\end{pmatrix} = \begin{pmatrix}
-1 + s - t \\
2 \\
s \\
t \\
\end{pmatrix}
\]  
(2.10)

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 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 (2.9). 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
2.3. EXERCISES

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 reduced echelon form rather than just echelon form.

In summary,

**Definition 2.2.18** A system of linear equations is a list of equations,

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

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

\[
\sum_{j=1}^{n} a_{ij}x_j = f_j, \quad i = 1, 2, 3, \ldots, m
\]

It is desired to find \((x_1, x_2, \ldots, x_n)\) 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.

2.3 Exercises

1. Find the point, \((x_1, y_1)\) which lies on both lines, \(x + 3y = 1\) and \(4x - y = 3\).
2. Solve Problem 1 graphically. That is, graph each line and see where they intersect.
3. Find the point of intersection of the two lines \(3x + y = 3\) and \(x + 2y = 1\).
4. Solve Problem 3 graphically. That is, graph each line and see where they intersect.
5. Do the three lines, \(x + 2y = 1\), \(2x - y = 1\), and \(4x + 3y = 3\) have a common point of intersection? If so, find the point and if not, tell why they don’t have such a common point of intersection.
6. Do the three planes, \(x + y - 3z = 2\), \(2x + y + z = 1\), and \(3x + 2y - 2z = 0\) have a common point of intersection? If so, find one and if not, tell why there is no such point.
7. You have a system of \(k\) equations in two variables, \(k \geq 2\). Explain the geometric significance of

   (a) No solution.
(b) A unique solution.
(c) An infinite number of solutions.

8. Here is an augmented matrix in which * denotes an arbitrary number and ■ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

\[
\begin{pmatrix}
■ & * & * & * & * & | & * \\
0 & ■ & * & 0 & | & * \\
0 & 0 & ■ & * & | & *
\end{pmatrix}
\]

9. Here is an augmented matrix in which * denotes an arbitrary number and ■ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

\[
\begin{pmatrix}
■ & * & | & * \\
0 & ■ & * & | & *
\end{pmatrix}
\]

10. Here is an augmented matrix in which * denotes an arbitrary number and ■ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

\[
\begin{pmatrix}
■ & * & * & * & | & * \\
0 & ■ & 0 & * & 0 & | & * \\
0 & 0 & 0 & ■ & 0 & | & *
\end{pmatrix}
\]

11. Here is an augmented matrix in which * denotes an arbitrary number and ■ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

\[
\begin{pmatrix}
■ & * & * & * & | & * \\
0 & ■ & * & 0 & * & | & *
\end{pmatrix}
\]

12. Find \( h \) such that

\[
\begin{pmatrix}
2 & h & | & 4 \\
3 & 6 & | & 7
\end{pmatrix}
\]

is the augmented matrix of an inconsistent matrix.

13. Find \( h \) such that

\[
\begin{pmatrix}
1 & h & | & 3 \\
2 & 4 & | & 6
\end{pmatrix}
\]

is the augmented matrix of a consistent matrix.

14. Find \( h \) such that

\[
\begin{pmatrix}
1 & 1 & | & 4 \\
3 & h & | & 12
\end{pmatrix}
\]

is the augmented matrix of a consistent matrix.
15. 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.

\[
\begin{pmatrix}
1 & h & 2 \\
2 & 4 & k
\end{pmatrix}
\]

16. 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.

\[
\begin{pmatrix}
1 & 2 & 2 \\
2 & h & k
\end{pmatrix}
\]

17. Determine if the system is consistent.

\[
\begin{align*}
x + 2y + z - w &= 2 \\
x - y + z + w &= 1 \\
2x + y - z &= 1 \\
4x + 2y + z &= 5
\end{align*}
\]

18. Determine if the system is consistent.

\[
\begin{align*}
x + 2y + z - w &= 2 \\
x - y + z + w &= 0 \\
2x + y - z &= 1 \\
4x + 2y + z &= 3
\end{align*}
\]

19. Find the general solution of the system whose augmented matrix is

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

20. Find the general solution of the system whose augmented matrix is

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

21. Find the general solution of the system whose augmented matrix is

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

22. Find the general solution of the system whose augmented matrix is

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

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

24. 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.

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

26. 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.

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

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

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

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

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

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

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

34. 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 people.

35. 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 \(z = \frac{1}{4}(10 + 0 + w + x).\)
36. 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?

37. Give the complete solution to the system of equations, $-9x + 15y = 66$, $-11x + 18y = 79$, $-x + y = 4$, and $z = 3$. 
Matrices

3.0.1 Outcomes

1. Perform the basic matrix operations of matrix addition, scalar multiplication, transposition and matrix multiplication. Identify when these operations are not defined. Represent the basic operations in terms of double subscript notation.

2. Recall and prove algebraic properties for matrix addition, scalar multiplication, transposition, and matrix multiplication. Apply these properties to manipulate an algebraic expression involving matrices.

3. Evaluate the inverse of a matrix using Gauss Jordan elimination.

4. Recall the cancellation laws for matrix multiplication. Demonstrate when cancellation laws do not apply.

5. Recall and prove identities involving matrix inverses.

3.1 Matrix Arithmetic

3.1.1 Addition And Scalar Multiplication Of 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 such rectangular arrays of numbers are important from many other different points of view. Numbers are also called scalars. In these notes numbers will always be either real or complex numbers.

A matrix is a rectangular array of numbers. Several of them are referred to as matrices. For example, here is a matrix.

\[
\begin{pmatrix}
1 & 2 & 3 & 4 \\
5 & 2 & 8 & 7 \\
6 & -9 & 1 & 2
\end{pmatrix}
\]

The size or dimension of a matrix is defined as \(m \times n\) where \(m\) is the number of rows and \(n\) is the number of columns. The above 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 \(\begin{pmatrix}1 \\ 5 \\ 6\end{pmatrix}\). When specifying the size of a matrix, you always list the number of rows before the number of columns. Also, you can remember the columns are like columns in a Greek temple. They stand upright while the rows just lay 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. The entry in the \(i^{th}\) row and the \(j^{th}\) column of this matrix is denoted by \(a_{ij}\). 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. Matrices can be added multiplied by a scalar, and multiplied by other matrices. To illustrate scalar multiplication, consider the following example in which a matrix is being multiplied by the scalar, 3.

\[
3 \begin{pmatrix}
1 & 2 & 3 & 4 \\
5 & 2 & 8 & 7 \\
6 & -9 & 1 & 2
\end{pmatrix} = \begin{pmatrix}
3 & 6 & 9 & 12 \\
15 & 6 & 24 & 21 \\
18 & -27 & 3 & 6
\end{pmatrix} .
\]

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 must be the same size to be added. The sum of two matrices is a matrix which is obtained by adding the 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

\[
\begin{pmatrix}
0 & 0 \\
0 & 0 \\
0 & 0
\end{pmatrix} \neq \begin{pmatrix}
0 & 0
\end{pmatrix}
\]

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 3.1.1** (Scalar Multiplication) If \(A = (a_{ij})\) and \(k\) is a scalar, then \(kA = (ka_{ij})\).

**Example 3.1.2** \(7 \begin{pmatrix}
2 & 0 \\
1 & -4
\end{pmatrix} = \begin{pmatrix}
14 & 0 \\
7 & -28
\end{pmatrix} .\)

**Definition 3.1.3** (Addition) If \(A = (a_{ij})\) and \(B = (b_{ij})\) are two \(m \times n\) matrices. Then \(A + B = C\) where

\[C = (c_{ij})\]

for \(c_{ij} = a_{ij} + b_{ij}\).

**Example 3.1.4**

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

To save on notation, we will often use \(A_{ij}\) to refer to the \(ij^{th}\) entry of the matrix, \(A\).

**Definition 3.1.5** (The zero matrix) The \(m \times n\) zero matrix is the \(m \times n\) matrix having every entry equal to zero. It is denoted by \(0\).

**Example 3.1.6** The \(2 \times 3\) zero matrix is \(\begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix} .\).
Note there are $2 \times 3$ zero matrices, $3 \times 4$ zero matrices, etc. In fact there is a zero matrix for every size.

**Definition 3.1.7 (Equality of matrices)** Let $A$ and $B$ be two matrices. Then $A = B$ means that the two matrices are of the same size and for $A = (a_{ij})$ and $B = (b_{ij})$, $a_{ij} = b_{ij}$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$.

The following properties of matrices can be easily verified. You should do so.

- **Commutative Law Of Addition.**
  
  $A + B = B + A,$  
  \hspace{1cm} (3.1)

- **Associative Law for Addition.**
  
  $(A + B) + C = A + (B + C),$  
  \hspace{1cm} (3.2)

- **Existence of an Additive Identity**
  
  $A + 0 = A,$  
  \hspace{1cm} (3.3)

- **Existence of an Additive Inverse**
  
  $A + (-A) = 0,$  
  \hspace{1cm} (3.4)

Also for $\alpha, \beta$ scalars, the following additional properties hold.

- **Distributive law over Matrix Addition.**
  
  $\alpha (A + B) = \alpha A + \alpha B,$  
  \hspace{1cm} (3.5)

- **Distributive law over Scalar Addition**
  
  $(\alpha + \beta) A = \alpha A + \beta A,$  
  \hspace{1cm} (3.6)

- **Associative law for Scalar Multiplication**
  
  $\alpha (\beta A) = \alpha \beta (A),$  
  \hspace{1cm} (3.7)

- **Rule for Multiplication by 1.**
  
  $1A = A.$  
  \hspace{1cm} (3.8)

As an example, consider the Commutative Law of Addition. Let $A + B = C$ and $B + A = D$. Why is $D = C$?

\[ C_{ij} = A_{ij} + B_{ij} = B_{ij} + A_{ij} = D_{ij}. \]

Therefore, $C = D$ because the $ij^{th}$ entries are the same. Note that the conclusion follows from the commutative law of addition of numbers.
3.1.2 Multiplication Of Matrices

Definition 3.1.8 Matrices which are $n \times 1$ or $1 \times n$ are called vectors and are often denoted by a bold letter. Thus the $n \times 1$ matrix

$$\mathbf{x} = \begin{pmatrix} x_1 \\
\vdots \\
x_n \end{pmatrix}$$

is also called a column vector. The $1 \times n$ matrix

$$(x_1 \cdots x_n)$$

is called a row vector.

Although the following description of matrix multiplication may seem strange, it is in fact the most important and useful of the matrix operations. To begin with consider the case where a matrix is multiplied by a column vector. We will illustrate the general definition by first considering a special case.

$$\begin{pmatrix} 1 & 2 & 3 \\
4 & 5 & 6 \end{pmatrix} \begin{pmatrix} 7 \\
8 \\
9 \end{pmatrix} = ?$$

One way to remember this is as follows. Slide the vector, placing it on top the two rows as shown and then do the indicated operation.

$$\begin{pmatrix} 7 & 8 & 9 \\
1 & 2 & 3 \\
4 & 5 & 6 \end{pmatrix} \rightarrow \begin{pmatrix} 7 \times 1 + 8 \times 2 + 9 \times 3 \\
7 \times 4 + 8 \times 5 + 9 \times 6 \end{pmatrix} = \begin{pmatrix} 50 \\
122 \end{pmatrix}.$$ 

multiply the numbers on the top by the numbers on the bottom and add them up to get a single number for each row of the matrix as shown above.

In more general terms,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\
a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} x_1 \\
x_2 \\
x_3 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \end{pmatrix}.$$ 

In general, here is the definition of how to multiply an $(m \times n)$ matrix times a $(n \times 1)$ matrix.

Definition 3.1.9 Let $A = A_{ij}$ be an $m \times n$ matrix and let $\mathbf{v}$ be an $n \times 1$ matrix,

$$\mathbf{v} = \begin{pmatrix} v_1 \\
\vdots \\
v_n \end{pmatrix}$$

Then $A\mathbf{v}$ is an $m \times 1$ matrix and the $i^{th}$ component of this matrix is

$$(A\mathbf{v})_i = A_{i1}v_1 + A_{i2}v_2 + \cdots + A_{in}v_n = \sum_{j=1}^{n} A_{ij}v_j.$$
Thus
\[ A v = \left( \begin{array}{c} \sum_{j=1}^{n} A_{1j}v_j \\ \vdots \\ \sum_{j=1}^{n} A_{mj}v_j \end{array} \right). \] (3.9)

In other words, if
\[ A = (a_1, \ldots, a_n) \]
where the \( a_k \) are the columns,
\[ A v = \sum_{k=1}^{n} v_k a_k \]
This follows from (3.9) and the observation that the \( j \)th column of \( A \) is
\[ \left( \begin{array}{c} A_{1j} \\ A_{2j} \\ \vdots \\ A_{mj} \end{array} \right) \]
so (3.9) reduces to
\[ v_1 \left( \begin{array}{c} A_{11} \\ A_{21} \\ \vdots \\ A_{m1} \end{array} \right) + v_2 \left( \begin{array}{c} A_{12} \\ A_{22} \\ \vdots \\ A_{m2} \end{array} \right) + \cdots + v_k \left( \begin{array}{c} A_{1k} \\ A_{2k} \\ \vdots \\ A_{mk} \end{array} \right) \]
Note also that multiplication by an \( m \times n \) matrix takes an \( n \times 1 \) matrix, and produces an \( m \times 1 \) matrix.

Here is another example.

Example 3.1.10 Compute
\[ \left( \begin{array}{cccc} 1 & 2 & 1 & 3 \\ 0 & 2 & 1 & -2 \\ 2 & 1 & 4 & 1 \end{array} \right) \left( \begin{array}{c} 1 \\ 2 \\ 0 \\ 1 \end{array} \right). \]

First of all this is of the form \((3 \times 4)(4 \times 1)\) and so the result should be a \((3 \times 1)\). Note how the inside numbers cancel. To get the element in the second row and first and only column, compute
\[ \sum_{k=1}^{4} a_{2k}v_k = a_{21}v_1 + a_{22}v_2 + a_{23}v_3 + a_{24}v_4 
= 0 \times 1 + 2 \times 2 + 1 \times 0 + (-2) \times 1 = 2. \]
You should do the rest of the problem and verify
\[ \left( \begin{array}{cccc} 1 & 2 & 1 & 3 \\ 0 & 2 & 1 & -2 \\ 2 & 1 & 4 & 1 \end{array} \right) \left( \begin{array}{c} 1 \\ 2 \\ 0 \\ 1 \end{array} \right) = \left( \begin{array}{c} 8 \\ 2 \\ 5 \end{array} \right). \]
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.

For \( A \) and \( B \) matrices, in order to form the product, \( AB \) the number of columns of \( A \) must equal the number of rows of \( B \).

\[
\text{these must match!} \\
(m \times \ n) (n \times p) = m \times p
\]

Note the two outside numbers give the size of the product. Remember:

**[If the two middle numbers don’t match, you can’t multiply the matrices]**

**Definition 3.1.11** When the number of columns of \( A \) equals the number of rows of \( B \) the two matrices are said to be **conformable** and the product, \( AB \) is obtained as follows. Let \( A \) be an \( m \times n \) matrix and let \( B \) be an \( n \times p \) matrix. Then \( B \) is of the form

\[
B = (b_1, \cdots, b_p)
\]

where \( b_k \) is an \( n \times 1 \) matrix or column vector. Then the \( m \times p \) matrix, \( AB \) is defined as follows:

\[
AB \equiv (Ab_1, \cdots, Ab_p) \tag{3.10}
\]

where \( Ab_k \) is an \( m \times 1 \) matrix or column vector which gives the \( k^{th} \) column of \( AB \).

**Example 3.1.12** 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, is it possible to multiply these matrices. According to the above discussion it should be a \( 2 \times 3 \) matrix of the form

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

You know how to multiply a matrix times a vector and so you do so to obtain each of the three columns. Thus

\[
\begin{pmatrix}
1 & 2 & 1 \\
0 & 2 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 2 & 0 \\
0 & 3 & 1 \\
-2 & 1 & 1
\end{pmatrix}
= \begin{pmatrix}
-1 & 9 & 3 \\
-2 & 7 & 3
\end{pmatrix}
\]

**Example 3.1.13** Multiply the following.

\[
\begin{pmatrix}
1 & 2 & 0 \\
0 & 3 & 1 \\
-2 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 2 & 1 \\
0 & 2 & 1
\end{pmatrix}
\]
3.1. MATRIX ARITHMETIC

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!

### 3.1.3 The \(ij\)th Entry 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

\[
\sum_{k=1}^{n} A_{ik} B_{kj}.
\]

(3.11)

This shows the following definition for matrix multiplication in terms of the \(ij\)th entries of the product coincides with Definition 3.1.11.

**Definition 3.1.14** 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

\[
(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.
\]

(3.12)

**Example 3.1.15** Multiply if possible

\[
\begin{pmatrix}
1 & 2 \\
3 & 1 \\
2 & 6
\end{pmatrix}
\begin{pmatrix}
2 & 3 & 1 \\
7 & 6 & 2
\end{pmatrix}.
\]

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, the two matrices are conformable and it is possible to do the multiplication. The result should be a \(3 \times 3\) matrix. The answer is of the form

\[
\begin{pmatrix}
1 & 2 \\
3 & 1 \\
2 & 6
\end{pmatrix}
\begin{pmatrix}
2 & 7 \\
3 & 1 \\
2 & 6
\end{pmatrix}, \begin{pmatrix}
1 & 2 \\
3 & 1 \\
2 & 6
\end{pmatrix}
\begin{pmatrix}
3 & 6 \\
3 & 1 \\
2 & 6
\end{pmatrix}, \begin{pmatrix}
1 & 2 \\
3 & 1 \\
2 & 6
\end{pmatrix}
\begin{pmatrix}
1 & 2 \\
3 & 1 \\
2 & 6
\end{pmatrix}
\]

where the commas separate the columns in the resulting product. Thus the above product equals

\[
\begin{pmatrix}
16 & 15 & 5 \\
13 & 15 & 5 \\
46 & 42 & 14
\end{pmatrix}.
\]
a 3 × 3 matrix as desired. In terms of the $ij^{th}$ entries and the above definition, the entry in the third row and second column of the product should equal

$$\sum_j a_{3j} b_{2j} = a_{31} b_{12} + a_{32} b_{22}$$

$$= 2 \times 3 + 6 \times 6 = 42.$$ 

You should try a few more such examples to verify the above definition in terms of the $ij^{th}$ entries works for other entries.

**Example 3.1.16** *Multiply if possible*

$$\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$ 

This is not possible because it is of the form $(3 \times 2) (3 \times 3)$ and the middle numbers don't match. In other words the two matrices are not conformable in the indicated order.

**Example 3.1.17** *Multiply if possible*

$$\begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix}.$$ 

This is possible because in this case it is of the form $(3 \times 3) (3 \times 2)$ and the middle numbers do match so the matrices are conformable. 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 3.1.18** *Multiply if possible*

$$\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 & 0 \end{pmatrix}.$$ 

In this case you are trying to do $(3 \times 1) (1 \times 4)$. The inside numbers match so you can do it. Verify

$$\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 2 & 4 & 2 & 0 \\ 1 & 2 & 1 & 0 \end{pmatrix}.$$ 

### 3.1.4 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 the two products be equal then?

**Example 3.1.19** *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

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix}.$$
3.1. MATRIX ARITHMETIC

You see these are not equal. Again you cannot conclude that $AB = BA$ for matrix multiplication even when multiplication is defined in both orders. However, there are some properties which do hold.

**Proposition 3.1.20** If all multiplications and additions make sense, the following hold for matrices, $A, B, C$ and $a, b$ scalars.

\[ A(aB + bC) = a(AB) + b(AC) \]  
\[ (B + C)A = BA + CA \]  
\[ A(BC) = (AB)C \]

**Proof:** Using Definition 3.1.14,

\[ (A(aB + bC))_{ij} = \sum_k A_{ik} (aB + bC)_{kj} \]
\[ = \sum_k A_{ik} (aB_{kj} + bC_{kj}) \]
\[ = a \sum_k A_{ik}B_{kj} + b \sum_k A_{ik}C_{kj} \]
\[ = a(AB)_{ij} + b(AC)_{ij} \]
\[ = (a(AB) + b(AC))_{ij}. \]

Thus $A(B + C) = AB + AC$ as claimed. Formula (3.14) is entirely similar.

Formula (3.15) is the associative law of multiplication. Using Definition 3.1.14,

\[ (A(BC))_{ij} = \sum_k A_{ik} (BC)_{kj} \]
\[ = \sum_k A_{ik} \sum_l B_{kl}C_{lj} \]
\[ = \sum_l (AB)_{il}C_{lj} \]
\[ = ((AB)C)_{ij}. \]

This proves (3.15).

### 3.1.5 The Transpose

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.

\[
\begin{pmatrix}
1 & 4 \\
3 & 1 \\
2 & 6
\end{pmatrix}^T = \begin{pmatrix}
1 & 3 & 2 \\
4 & 1 & 6
\end{pmatrix}
\]

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. Here is the definition.

**Definition 3.1.21** Let $A$ be an $m \times n$ matrix. Then $A^T$ denotes the $n \times m$ matrix which is defined as follows.

\[ (A^T)_{ij} = A_{ji} \]
Example 3.1.22

\[
\begin{pmatrix}
1 & 2 & -6 \\
3 & 5 & 4
\end{pmatrix}^T = 
\begin{pmatrix}
1 & 2 \\
3 & 5 \\
-6 & 4
\end{pmatrix}.
\]

The transpose of a matrix has the following important properties.

Lemma 3.1.23 Let \( A \) be an \( m \times n \) matrix and let \( B \) be a \( n \times p \) matrix. Then

\[
(AB)^T = B^T A^T
\]

and if \( \alpha \) and \( \beta \) are scalars,

\[
(\alpha A + \beta B)^T = \alpha A^T + \beta B^T
\]

Proof: From the definition,

\[
\left((AB)^T\right)_{ij} = (AB)_{ji} = \sum_k A_{jk} B_{ki} = \sum_k \left(B^T\right)_{ik} \left(A^T\right)_{kj} = \left(B^T A^T\right)_{ij}
\]

The proof of Formula (3.17) is left as an exercise and this proves the lemma.

Definition 3.1.24 An \( n \times n \) matrix, \( A \) is said to be symmetric if \( A = A^T \). It is said to be skew symmetric if \( A = -A^T \).

Example 3.1.25 Let

\[
A = 
\begin{pmatrix}
2 & 1 & 3 \\
1 & 5 & -3 \\
3 & -3 & 7
\end{pmatrix}
\]

Then \( A \) is symmetric.

Example 3.1.26 Let

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

Then \( A \) is skew symmetric.

3.1.6 The Identity And Inverses

There is a special matrix called \( I \) and referred to as the identity matrix. It is always a square matrix, meaning the number of rows equals the number of columns and it has the property that there are ones down the main diagonal and zeroes elsewhere. Here are some identity matrices of various sizes.

\[
(1), 
\begin{pmatrix}
1 & 0 \\
0 & 1
\end{pmatrix}, 
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}, 
\begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\]
3.1. MATRIX ARITHMETIC

The first is the $1 \times 1$ identity matrix, the second is the $2 \times 2$ identity matrix, the third is the $3 \times 3$ identity matrix, and the fourth is the $4 \times 4$ identity matrix. By extension, you can likely see what the $n \times n$ identity matrix would be. It is so important that there is a special symbol to denote the $i_{th}$ entry of the identity matrix

$$I_{ij} = \delta_{ij}$$

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 3.1.27** 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 3.1.28** An $n \times n$ matrix, $A$ has an inverse, $A^{-1}$ if and only if $AA^{-1} = A^{-1}A = I$. Such a matrix is called invertible.

It is very important to observe that the inverse of a matrix, if it exists, is unique. Another way to think of this is that if it acts like the inverse, then it is the inverse.

**Theorem 3.1.29** Suppose $A^{-1}$ exists and $AB = BA = I$. Then $B = A^{-1}$.

**Proof:**

$$A^{-1} = A^{-1}I = A^{-1}(AB) = (A^{-1}A)B = IB = B.$$ 

Unlike ordinary multiplication of numbers, it can happen that $A \neq 0$ but $A$ may fail to have an inverse. This is illustrated in the following example.

**Example 3.1.30** 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,

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and if $A^{-1}$ existed, this could not happen because you could write

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = A^{-1}\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = A^{-1}\begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix} = (A^{-1}A)\begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix} = I\begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix},$$

a contradiction. Thus the answer is that $A$ does not have an inverse.
Example 3.1.31 Let \( A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \). Show \( \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \) is the inverse of \( A \).

To check this, multiply
\[
\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}
\]
and
\[
\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}
\]
showing that this matrix is indeed the inverse of \( A \).

3.1.7 Finding The Inverse Of A Matrix

In the last example, how would you find \( A^{-1} \)? You wish to find a matrix, \( \begin{pmatrix} x & z \\ y & w \end{pmatrix} \) such that
\[
\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x & z \\ y & w \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}
\]
This requires the solution of the systems of equations,
\[x + y = 1, x + 2y = 0\]
and
\[z + w = 0, z + 2w = 1.\]
Writing the augmented matrix for these two systems gives
\[
\begin{pmatrix} 1 & 1 & | & 1 \\ 1 & 2 & | & 0 \end{pmatrix}
\]
for the first system and
\[
\begin{pmatrix} 1 & 1 & | & 0 \\ 1 & 2 & | & 1 \end{pmatrix}
\]
for the second. Let’s solve the first system. Take \((-1)\) times the first row and add to the second to get
\[
\begin{pmatrix} 1 & 1 & | & 1 \\ 0 & 1 & | & -1 \end{pmatrix}
\]
Now take \((-1)\) times the second row and add to the first to get
\[
\begin{pmatrix} 1 & 0 & | & 2 \\ 0 & 1 & | & -1 \end{pmatrix}
\]
Putting in the variables, this says \( x = 2 \) and \( y = -1 \).

Now solve the second system, (3.19) to find \( z \) and \( w \). Take \((-1)\) times the first row and add to the second to get
\[
\begin{pmatrix} 1 & 1 & | & 0 \\ 0 & 1 & | & 1 \end{pmatrix}
\]
Now take \((-1)\) times the second row and add to the first to get
\[
\begin{pmatrix} 1 & 0 & | & -1 \\ 0 & 1 & | & 1 \end{pmatrix}
\]
Putting in the variables, this says $z = -1$ and $w = 1$. Therefore, the inverse is

$$
\begin{bmatrix}
2 & -1 \\
-1 & 1
\end{bmatrix}.
$$

Didn’t the above seem rather repetitive? Note that exactly the same row operations were used in both systems. In each case, the end result was something of the form $(I|v)$ where $I$ is the identity and $v$ gave a column of the inverse. In the above, \( \begin{bmatrix} x \\ y \end{bmatrix} \), the first column of the inverse was obtained first and then the second column \( \begin{bmatrix} z \\ w \end{bmatrix} \).

To simplify this procedure, you could have written

$$
\begin{bmatrix}
1 & 1 \\
1 & 2
\end{bmatrix} | \begin{bmatrix} 1 \\ 0 \end{bmatrix}
$$

and row reduced till you obtained

$$
\begin{bmatrix}
1 & 0 & 2 & -1 \\
0 & 1 & -1 & 1
\end{bmatrix}
$$

and read off the inverse as the $2 \times 2$ matrix on the right side.

This is the reason for the following simple procedure for finding the inverse of a matrix. This procedure is called the **Gauss-Jordan procedure**.

**Procedure 3.1.32** 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, if possible do row operations until you obtain an $n \times 2n$ matrix of the form

$$
(I|B). \quad (3.20)
$$

When this has been done, $B = A^{-1}$. If it is impossible to row reduce to a matrix of the form $(I|B)$, then $A$ has no inverse.

**Example 3.1.33** Let $A = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 3 & 1 & -1 \end{bmatrix}$. Find $A^{-1}$ if it exists.

Set up the augmented matrix, $(A|I)$

$$
\begin{bmatrix}
1 & 2 & 2 & 1 & 0 & 0 \\
1 & 0 & 2 & 0 & 1 & 0 \\
3 & 1 & -1 & 0 & 0 & 1
\end{bmatrix}
$$

Next take $(−1)$ times the first row and add to the second followed by $(−3)$ times the first row added to the last. This yields

$$
\begin{bmatrix}
1 & 2 & 2 & 1 & 0 & 0 \\
0 & -2 & -2 & -1 & 1 & 0 \\
0 & -5 & -7 & -3 & 0 & 1
\end{bmatrix}.
$$

Then take 5 times the second row and add to -2 times the last row.

$$
\begin{bmatrix}
1 & 2 & 2 & 1 & 0 & 0 \\
0 & -10 & -10 & -5 & 5 & 0 \\
0 & 0 & 14 & 1 & 5 & -2
\end{bmatrix}
$$
Next take the last row and add to \((-7)\) times the top row. This yields
\[
\begin{pmatrix}
-7 & -14 & 0 & | & -6 & 5 & -2 \\
0 & -10 & 0 & | & -5 & 5 & 0 \\
0 & 0 & 14 & | & 1 & 5 & -2
\end{pmatrix}.
\]

Now take \((-7/5)\) times the second row and add to the top.
\[
\begin{pmatrix}
-7 & 0 & 0 & | & 1 & -2 & -2 \\
0 & -10 & 0 & | & -5 & 5 & 0 \\
0 & 0 & 14 & | & 1 & 5 & -2
\end{pmatrix}.
\]

Finally divide the top row by -7, the second row by -10 and the bottom row by 14 which yields
\[
\begin{pmatrix}
1 & 0 & 0 & | & -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\
0 & 1 & 0 & | & \frac{1}{7} & -\frac{1}{7} & 0 \\
0 & 0 & 1 & | & \frac{1}{14} & \frac{5}{14} & -\frac{1}{7}
\end{pmatrix}.
\]

Therefore, the inverse is
\[
\begin{pmatrix}
-\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\
\frac{1}{7} & -\frac{1}{7} & 0 \\
\frac{1}{14} & \frac{5}{14} & -\frac{1}{7}
\end{pmatrix}.
\]

**Example 3.1.34** Let \(A = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 2 & 2 & 4 \end{pmatrix} \). Find \(A^{-1}\) if it exists.

Write the augmented matrix, \((A|I)\)
\[
\begin{pmatrix}
1 & 2 & 2 & | & 1 & 0 & 0 \\
1 & 0 & 2 & | & 0 & 1 & 0 \\
2 & 2 & 4 & | & 0 & 0 & 1
\end{pmatrix}
\]

and proceed to do row operations attempting to obtain \((I|A^{-1})\). Take \((-1)\) times the top row and add to the second. Then take \((-2)\) times the top row and add to the bottom.
\[
\begin{pmatrix}
1 & 2 & 2 & | & 1 & 0 & 0 \\
0 & -2 & 0 & | & -1 & 1 & 0 \\
0 & -2 & 0 & | & -2 & 0 & 1
\end{pmatrix}
\]

Next add \((-1)\) times the second row to the bottom row.
\[
\begin{pmatrix}
1 & 2 & 2 & | & 1 & 0 & 0 \\
0 & -2 & 0 & | & -1 & 1 & 0 \\
0 & 0 & 0 & | & -1 & -1 & 1
\end{pmatrix}
\]

At this point, you can see there will be no inverse because you have obtained a row of zeros in the left half of the augmented matrix, \((A|I)\). Thus there will be no way to obtain \(I\) on the left.
Example 3.1.35 Let \( A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix} \). Find \( A^{-1} \) if it exists.

Form the augmented matrix,
\[
\begin{pmatrix}
1 & 0 & 1 & | & 1 & 0 & 0 \\
1 & -1 & 1 & | & 0 & 1 & 0 \\
1 & 1 & -1 & | & 0 & 0 & 1
\end{pmatrix}
\]

Now do row operations until the \( n \times n \) matrix on the left becomes the identity matrix. This yields after some computations,
\[
\begin{pmatrix}
1 & 0 & 0 & | & 0 & \frac{1}{2} & \frac{1}{2} \\
0 & 1 & 0 & | & 1 & -1 & 0 \\
0 & 0 & 1 & | & 1 & -\frac{1}{2} & -\frac{1}{2}
\end{pmatrix}
\]

and so the inverse of \( A \) is the matrix on the right,
\[
\begin{pmatrix}
0 & \frac{1}{2} & \frac{1}{2} \\
1 & -1 & 0 \\
1 & -\frac{1}{2} & -\frac{1}{2}
\end{pmatrix}
\]

Checking the answer is easy. Just multiply the matrices and see if it works.
\[
\begin{pmatrix}
1 & 0 & 1 \\
1 & -1 & 1 \\
1 & 1 & -1
\end{pmatrix}
\begin{pmatrix}
0 & \frac{1}{2} & \frac{1}{2} \\
1 & -1 & 0 \\
1 & -\frac{1}{2} & -\frac{1}{2}
\end{pmatrix}
= \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]

Always check your answer because if you are like some of us, you will usually have made a mistake.

3.2 Exercises

1. Here are some matrices:

\[
A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 7 \end{pmatrix}, B = \begin{pmatrix} 3 & -1 & 2 \\ -3 & 2 & 1 \end{pmatrix}, C = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}, D = \begin{pmatrix} -1 & 2 \\ 2 & -3 \end{pmatrix}, E = \begin{pmatrix} 2 \\ 3 \end{pmatrix}
\]

Find if possible \(-3A, 3B - A, AC, CB, AE, EA\). If it is not possible explain why.

2. Here are some matrices:

\[
A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix}, C = \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 & 1 \\ 4 & -3 \end{pmatrix}, E = \begin{pmatrix} 1 \\ 3 \end{pmatrix}
\]

Find if possible \(-3A, 3B - A, AC, CA, AE, EA, BE, DE\). If it is not possible explain why.
3. Here are some matrices:

\[ A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \]

\[ C = \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 & 1 \\ 4 & -3 \end{pmatrix}, E = \begin{pmatrix} 1 \\ 3 \end{pmatrix}. \]

Find if possible \(-3A^T, 3B - A^T, AC, CA, AE, E^TB, DE, EE^T, E^T E\). If it is not possible explain why.

4. Here are some matrices:

\[ A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \]

\[ C = \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 & 1 \\ 4 & -3 \end{pmatrix}, E = \begin{pmatrix} 1 \\ 3 \end{pmatrix}. \]

Find the following if possible and explain why it is not possible if this is the case.

\[ AD, DA, D^T B, D^T E, E^T D, DE, EE^T. \]

5. 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.

(a) \( AB \)
(b) \( BA \)
(c) \( AC \)
(d) \( CA \)
(e) \( CB \)
(f) \( BC \)

6. Let \( A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \), \( B = \begin{pmatrix} 1 & 2 \\ 3 & k \end{pmatrix} \). Is it possible to choose \( k \) such that \( AB = BA \)? If so, what should \( k \) equal?

7. Let \( A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \), \( B = \begin{pmatrix} 1 & 2 \\ 1 & k \end{pmatrix} \). Is it possible to choose \( k \) such that \( AB = BA \)? If so, what should \( k \) equal?

8. Let \( x = (-1, -1, 1) \) and \( y = (0, 1, 2) \). Find \( x^T y \) and \( xy^T \) if possible.

9. Write \( \begin{pmatrix} x_1 - x_2 + 2x_3 \\ 2x_3 + x_1 \\ 3x_3 \\ 3x_4 + 3x_2 + x_1 \end{pmatrix} \) in the form \( A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \) where \( A \) is an appropriate matrix.

10. Suppose \( A \) and \( B \) are square matrices of the same size. Which of the following are correct?

(a) \( (A - B)^2 = A^2 - 2AB + B^2 \)
3.2. EXERCISES

(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) \((A + B)(A - B) = A^2 - B^2\)

11. Let \(A = \begin{pmatrix} -1 & -1 \\ 3 & 3 \end{pmatrix}\). Find all \(2 \times 2\) matrices, \(B\) such that \(AB = 0\).

12. In (3.1) - (3.8) describe \(-A\) and 0.

13. Let \(A\) be an \(n \times n\) matrix. Show \(A\) equals the sum of a symmetric and a skew symmetric matrix.

14. 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.

15. Using only the properties (3.1) - (3.8) show \(-A\) is unique.

16. Using only the properties (3.1) - (3.8) show 0 is unique.

17. Using only the properties (3.1) - (3.8) show 0\(A = 0\). Here the 0 on the left is the scalar 0 and the 0 on the right is the zero for \(m \times n\) matrices.

18. Using only the properties (3.1) - (3.8) and previous problems show \((-1)A = -A\).


20. Prove that \(I_mA = A\) where \(A\) is an \(m \times n\) matrix.

21. 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.

22. 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.

23. 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.
24. 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.

25. Give an example of matrices, \( A, B, C \) such that \( B \neq C, A \neq 0 \), and yet \( AB = AC \).

26. Suppose \( AB = AC \) and \( A \) is an invertible \( n \times n \) matrix. Does it follow that \( B = C \)? Explain why or why not. What if \( A \) were a non invertible \( n \times n \) matrix?

27. Find your own examples:
   (a) \( 2 \times 2 \) matrices, \( A \) and \( B \) such that \( A \neq 0, B \neq 0 \) with \( AB \neq BA \).
   (b) \( 2 \times 2 \) matrices, \( A \) and \( B \) such that \( A \neq 0, B \neq 0 \), but \( AB = 0 \).
   (c) \( 2 \times 2 \) matrices, \( A, D, \) and \( C \) such that \( A \neq 0, C \neq D \), but \( AC = AD \).

28. Explain why if \( AB = AC \) and \( A^{-1} \) exists, then \( B = C \).

29. Give an example of a matrix, \( A \) such that \( A^2 = I \) and yet \( A \neq I \) and \( A \neq -I \).

30. Give an example of matrices, \( A, B \) such that neither \( A \) nor \( B \) equals zero and yet \( AB = 0 \).

31. Give another example other than the one given in this section of two square matrices, \( A \) and \( B \) such that \( AB \neq BA \).

32. Show that if \( A^{-1} \) exists for an \( n \times n \) matrix, then it is unique. That is, if \( BA = I \) and \( AB = I \), then \( B = A^{-1} \).

33. Show \( (AB)^{-1} = B^{-1}A^{-1} \).

34. Show that if \( A \) is an invertible \( n \times n \) matrix, then so is \( A^T \) and \( (A^T)^{-1} = (A^{-1})^T \).

35. 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 \( x = A^{-1}b \).

36. Prove that if \( A^{-1} \) exists and \( Ax = 0 \) then \( x = 0 \).

37. Show that \( (ABC)^{-1} = C^{-1}B^{-1}A^{-1} \) by verifying that \( (ABC)(C^{-1}B^{-1}A^{-1}) = I \).
   Assume for now that if a matrix acts like the inverse on one side of a matrix, then it is the inverse and will work as such on the other side.
Determinants

4.0.1 Outcomes
1. Evaluate the determinant of a square matrix using by applying
   (a) the cofactor formula or
   (b) row operations.
2. Recall the general properties of determinants.
3. Recall that the determinant of a product of matrices is the product of the determinants.
   Recall that the determinant of a matrix is equal to the determinant of its transpose.
4. Apply Cramer’s Rule to solve a $2 \times 2$ or a $3 \times 3$ linear system.
5. Use determinants to determine whether a matrix has an inverse.
6. Evaluate the inverse of a matrix using cofactors.

4.1 Basic Techniques And Properties

4.1.1 Cofactors And $2 \times 2$ Determinants
Let $A$ be an $n \times n$ matrix. The determinant of $A$, denoted as $\det (A)$ is a number. If the matrix is a $2 \times 2$ matrix, this number is very easy to find.

**Definition 4.1.1** Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then
\[
\det (A) \equiv ad - cb.
\]

The determinant is also often denoted by enclosing the matrix with two vertical lines. Thus
\[
\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \left| \begin{array}{cc} a & b \\ c & d \end{array} \right|.
\]

**Example 4.1.2** Find $\det \begin{pmatrix} 2 & 4 \\ -1 & 6 \end{pmatrix}$.

From the definition this is just $(2)(6) - (-1)(4) = 16$.

Having defined what is meant by the determinant of a $2 \times 2$ matrix, what about a $3 \times 3$ matrix?
Definition 4.1.3 Suppose $A$ is a $3 \times 3$ matrix. The $ij^{th}$ minor, denoted as $\text{minor}(A)_{ij}$, is the determinant of the $2 \times 2$ matrix which results from deleting the $i^{th}$ row and the $j^{th}$ column.

Example 4.1.4 Consider the matrix,

$$
\begin{pmatrix}
1 & 2 & 3 \\
4 & 3 & 2 \\
3 & 2 & 1
\end{pmatrix}.
$$

The $(1, 2)$ minor is the determinant of the $2 \times 2$ matrix which results when you delete the first row and the second column. This minor is therefore

$$
\det \begin{pmatrix} 4 & 2 \\ 3 & 1 \end{pmatrix} = -2.
$$

The $(2, 3)$ minor is the determinant of the $2 \times 2$ matrix which results when you delete the second row and the third column. This minor is therefore

$$
\det \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} = -4.
$$

Definition 4.1.5 Suppose $A$ is a $3 \times 3$ matrix. The $ij^{th}$ cofactor is defined to be $(-1)^{i+j} \times (ij^{th} \text{ minor})$. In words, you multiply $(-1)^{i+j}$ times the $ij^{th}$ minor to get the $ij^{th}$ cofactor. The cofactors of a matrix are so important that special notation is appropriate when referring to them. The $ij^{th}$ cofactor of a matrix, $A$ will be denoted by $\text{cof} (A)_{ij}$. It is also convenient to refer to the cofactor of an entry of a matrix as follows. For $a_{ij}$ an entry of the matrix, its cofactor is just $\text{cof} (A)_{ij}$. Thus the cofactor of the $ij^{th}$ entry is just the $ij^{th}$ cofactor.

Example 4.1.6 Consider the matrix,

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

The $(1, 2)$ minor is the determinant of the $2 \times 2$ matrix which results when you delete the first row and the second column. This minor is therefore

$$
\det \begin{pmatrix} 4 & 2 \\ 3 & 1 \end{pmatrix} = -2.
$$

It follows

$$
\text{cof} (A)_{12} = (-1)^{1+2} \det \begin{pmatrix} 4 & 2 \\ 3 & 1 \end{pmatrix} = (-1)^{1+2} (-2) = 2
$$

The $(2, 3)$ minor is the determinant of the $2 \times 2$ matrix which results when you delete the second row and the third column. This minor is therefore

$$
\det \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} = -4.
$$

Therefore,

$$
\text{cof} (A)_{23} = (-1)^{2+3} \det \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} = (-1)^{2+3} (-4) = 4.
$$

Similarly,

$$
\text{cof} (A)_{22} = (-1)^{2+2} \det \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix} = -8.
$$
Definition 4.1.7 The determinant of a $3 \times 3$ matrix, $A$, is obtained by picking a row (column) and taking the product of each entry in that row (column) with its cofactor and adding these up. This process when applied to the $i^{th}$ row (column) is known as expanding the determinant along the $i^{th}$ row (column).

Example 4.1.8 Find the determinant of

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

Here is how it is done by “expanding along the first column”.

$$ \begin{align*} 
&= 1(-1)^{1+1} \begin{vmatrix} 3 & 2 \\ 2 & 1 \end{vmatrix} + 4(-1)^{2+1} \begin{vmatrix} 2 & 3 \\ 2 & 1 \end{vmatrix} + 3(-1)^{3+1} \begin{vmatrix} 2 & 3 \\ 3 & 2 \end{vmatrix} \\
&= 0. 
\end{align*}$$ 

You see, we just followed the rule in the above definition. We took the 1 in the first column and multiplied it by its cofactor, the 4 in the first column and multiplied it by its cofactor, and the 3 in the first column and multiplied it by its cofactor. Then we added these numbers together.

You could also expand the determinant along the second row as follows.

$$\begin{align*} 
&= 4(-1)^{2+1} \begin{vmatrix} 2 & 3 \\ 2 & 1 \end{vmatrix} + 3(-1)^{3+2} \begin{vmatrix} 1 & 3 \\ 3 & 1 \end{vmatrix} + 2(-1)^{4+3} \begin{vmatrix} 1 & 2 \\ 3 & 2 \end{vmatrix} \\
&= 0. 
\end{align*}$$ 

Observe this gives the same number. You should try expanding along other rows and columns. If you don’t make any mistakes, you will always get the same answer.

What about a $4 \times 4$ matrix? You know how to find the determinant of a $3 \times 3$ matrix. The pattern is the same.

Definition 4.1.9 Suppose $A$ is a $4 \times 4$ matrix. The $ij^{th}$ minor is the determinant of the $3 \times 3$ matrix you obtain when you delete the $i^{th}$ row and the $j^{th}$ column. The $ij^{th}$ cofactor, $\text{cof}(A)_{ij}$, is defined to be $(-1)^{i+j}$ times the $ij^{th}$ minor to get the $ij^{th}$ cofactor.

Definition 4.1.10 The determinant of a $4 \times 4$ matrix, $A$, is obtained by picking a row (column) and taking the product of each entry in that row (column) with its cofactor and adding these up. This process when applied to the $i^{th}$ row (column) is known as expanding the determinant along the $i^{th}$ row (column).

Example 4.1.11 Find $\det(A)$ where

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 4 & 2 & 3 \\ 1 & 3 & 4 & 5 \\ 3 & 4 & 3 & 2 \end{pmatrix}.$$

As in the case of a $3 \times 3$ matrix, you can expand this along any row or column. Lets pick the third column. $\det(A) =$

$$3(-1)^{1+3} \begin{vmatrix} 5 & 4 & 3 \\ 1 & 3 & 5 \\ 3 & 4 & 2 \end{vmatrix} + 2(-1)^{2+3} \begin{vmatrix} 1 & 2 & 4 \\ 1 & 3 & 5 \\ 3 & 4 & 2 \end{vmatrix} +$$
DETERMINANTS

Now you know how to expand each of these 3×3 matrices along a row or a column. If you do so, you will get −12 assuming you make no mistakes. You could expand this matrix along any row or any column and assuming you make no mistakes, you will always get the same thing which is defined to be the determinant of the matrix, A. This method of evaluating a determinant by expanding along a row or a column is called the method of Laplace expansion.

Note that each of the four terms above involves three terms consisting of determinants of 2×2 matrices and each of these will need 2 terms. Therefore, there will be 4×3×2 = 24 terms to evaluate in order to find the determinant using the method of Laplace expansion. Suppose now you have a 10×10 matrix and you follow the above pattern for evaluating determinants. By analogy to the above, there will be 10! = 3,628,800 terms involved in the evaluation of such a determinant by Laplace expansion along a row or column. This is a lot of terms.

In addition to the difficulties just discussed, you should regard the above claim that you always get the same answer by picking any row or column with considerable skepticism. It is incredible and not at all obvious. However, it requires a little effort to establish it. This is done in the section on the theory of the determinant. The above examples motivate the following definitions, the second of which is incredible.

Definition 4.1.12 Let A = (a_{ij}) be an n×n matrix and suppose the determinant of a (n−1)×(n−1) matrix has been defined. Then a new matrix called the cofactor matrix, cof (A) is defined by 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)×(n−1) matrix which results, (This is called the ij-th minor of A.) and then multiply this number by (−1)^{i+j}. Thus (−1)^{i+j} × (the ij-th minor) equals the ij-th cofactor. To make the formulas easier to remember, cof (A)_{ij} will denote the ij-th entry of the cofactor matrix.

With this definition of the cofactor matrix, here is how to define the determinant of an n×n matrix.

Definition 4.1.13 Let A be an n×n matrix where n ≥ 2 and suppose the determinant of an (n−1)×(n−1) has been defined. Then

\[
det (A) = \sum_{j=1}^{n} a_{ij} \text{cof} (A)_{ij} = \sum_{i=1}^{n} a_{ij} \text{cof} (A)_{ij}.
\] (4.1)

The first formula consists of expanding the determinant along the i-th row and the second expands the determinant along the j-th column. This is called the method of Laplace expansion.

Theorem 4.1.14 Expanding the n×n matrix along any row or column always gives the same answer so the above definition is a good definition.

4.1.2 The Determinant Of A Triangular Matrix

Notwithstanding the difficulties involved in using the method of Laplace expansion, certain types of matrices are very easy to deal with.
4.1. BASIC TECHNIQUES AND PROPERTIES

Definition 4.1.15 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.

\[
\begin{pmatrix}
\ast & \ast & \cdots & \ast \\
0 & \ast & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ast \\
0 & \cdots & 0 & \ast
\end{pmatrix}
\]

A lower triangular matrix is defined similarly as a matrix for which all entries above the main diagonal are equal to zero.

You should verify the following using the above theorem on Laplace expansion.

Corollary 4.1.16 Let \( M \) be an upper (lower) triangular matrix. Then \( \det(M) \) is obtained by taking the product of the entries on the main diagonal.

Example 4.1.17 Let

\[
A = \begin{pmatrix}
1 & 2 & 3 & 77 \\
0 & 2 & 6 & 7 \\
0 & 0 & 3 & 33.7 \\
0 & 0 & 0 & -1
\end{pmatrix}
\]

Find \( \det(A) \).

From the above corollary, it suffices to take the product of the diagonal elements. Thus \( \det(A) = 1 \times 2 \times 3 \times (-1) = -6 \). Without using the corollary, you could expand along the first column. This gives

\[
\begin{vmatrix}
1 & 2 & 6 & 7 \\
0 & 0 & 3 & 33.7 \\
0 & 0 & -1 & 0
\end{vmatrix} + 0 \begin{vmatrix}
2 & 3 & 77 \\
0 & 0 & 3 & 33.7 \\
0 & 0 & 0 & -1
\end{vmatrix} + 0 \begin{vmatrix}
2 & 6 & 7 \\
0 & 0 & 3 & 33.7 \\
0 & -1 & 0
\end{vmatrix}
\]

and the only nonzero term in the expansion is

\[
\begin{vmatrix}
1 & 2 & 6 & 7 \\
0 & 0 & 3 & 33.7 \\
0 & 0 & -1 & 0
\end{vmatrix}.
\]

Now expand this along the first column to obtain

\[
1 \times \left( 2 \times \begin{vmatrix}
3 & 33.7 \\
0 & -1
\end{vmatrix} + 0 \begin{vmatrix}
6 & 7 \\
0 & -1
\end{vmatrix} + 0 \begin{vmatrix}
6 & 3 & 33.7 \\
3 & 3 & -1
\end{vmatrix} \right)
\]

\[
= 1 \times 2 \times \begin{vmatrix}
3 & 33.7 \\
0 & -1
\end{vmatrix}
\]

Next expand this last determinant along the first column to obtain the above equals

\[
1 \times 2 \times 3 \times (-1) = -6
\]

which is just the product of the entries down the main diagonal of the original matrix.
4.1.3 Properties Of Determinants

There are many properties satisfied by determinants. Some of these properties have to do with row operations. Recall the row operations.

**Definition 4.1.18** 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 itself.**

**Theorem 4.1.19** Let $A$ be an $n \times n$ matrix and let $A_1$ be a matrix which results from multiplying some row of $A$ by a scalar, $c$. Then $c \det(A) = \det(A_1)$.

**Example 4.1.20** Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $A_1 = \begin{pmatrix} 2 & 4 \\ 3 & 4 \end{pmatrix}$. Then $\det(A) = -2$, $\det(A_1) = -4$.

**Theorem 4.1.21** Let $A$ be an $n \times n$ matrix and let $A_1$ be a matrix which results from switching two rows of $A$. Then $\det(A) = -\det(A_1)$. Also, if one row of $A$ is a multiple of another row of $A$, then $\det(A) = 0$.

**Example 4.1.22** Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and let $A_1 = \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix}$. Then $\det(A) = -2$, $\det(A_1) = 2$.

**Theorem 4.1.23** Let $A$ be an $n \times n$ matrix and let $A_1$ be a matrix which results from applying row operation 3. That is you replace some row by a multiple of another row added to itself. Then $\det(A) = \det(A_1)$.

**Example 4.1.24** Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and let $A_1 = \begin{pmatrix} 1 & 2 \\ 4 & 6 \end{pmatrix}$. Then the second row of $A_1$ is one times the first row added to the second row. $\det(A) = -2$ and $\det(A_1) = -2$.

**Theorem 4.1.25** In Theorems 4.1.19 - 4.1.23 you can replace the word, “row” with the word “column”.

There are two other major properties of determinants which do not involve row operations.

**Theorem 4.1.26** Let $A$ and $B$ be two $n \times n$ matrices. Then

$$\det(AB) = \det(A)\det(B).$$

Also,

$$\det(A) = \det(A^T).$$

**Example 4.1.27** Compare $\det(AB)$ and $\det(A)\det(B)$ for

$$A = \begin{pmatrix} 1 & 2 \\ -3 & 2 \end{pmatrix}, B = \begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix}.$$
First
\[ AB = \begin{pmatrix} 1 & 2 \\ -3 & 2 \end{pmatrix} \begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} 11 & 4 \\ -1 & -4 \end{pmatrix} \]
and so
\[ \det(AB) = \det\begin{pmatrix} 11 & 4 \\ -1 & -4 \end{pmatrix} = -40. \]
Now
\[ \det(A) = \det\begin{pmatrix} 1 & 2 \\ -3 & 2 \end{pmatrix} = 8 \]
and
\[ \det(B) = \det\begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix} = -5. \]
Thus \( \det(A) \det(B) = 8 \times (-5) = -40. \)

4.1.4 Finding Determinants Using Row Operations

Theorems 4.1.23 - 4.1.25 can be used to find determinants using row operations. As pointed out above, the method of Laplace expansion will not be practical for any matrix of large size. Here is an example in which all the row operations are used.

Example 4.1.28 Find the determinant of the matrix,

\[ A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 1 & 2 & 3 \\ 4 & 5 & 4 & 3 \\ 2 & 2 & -4 & 5 \end{pmatrix} \]

Replace the second row by \((-5)\) times the first row added to it. Then replace the third row by \((-4)\) times the first row added to it. Finally, replace the fourth row by \((-2)\) times the first row added to it. This yields the matrix,

\[ B = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -9 & -13 & -17 \\ 0 & -3 & -8 & -13 \\ 0 & -2 & -10 & -3 \end{pmatrix} \]

and from Theorem 4.1.23, it has the same determinant as \( A \). Now using other row operations, \( \det(B) = \left(\frac{-1}{3}\right) \det(C) \) where

\[ C = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 11 & 22 \\ 0 & -3 & -8 & -13 \\ 0 & 6 & 30 & 9 \end{pmatrix} \]

The second row was replaced by \((-3)\) times the third row added to the second row. By Theorem 4.1.23 this didn’t change the value of the determinant. Then the last row was multiplied by \((-2)\). By Theorem 4.1.19 the resulting matrix has a determinant which is \((-3)\) times the determinant of the unmultiplied matrix. Therefore, we multiplied by \(-1/3\) to retain the correct value. Now replace the last row with 2 times the third added to it. This does not change the value of the determinant by Theorem 4.1.23. Finally switch
the third and second rows. This causes the determinant to be multiplied by \((-1)\). Thus
\[ \det(C) = -\det(D) \]
where
\[ D = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -3 & -8 & -13 \\ 0 & 0 & 11 & 22 \\ 0 & 0 & 14 & -17 \end{pmatrix} \]
You could do more row operations or you could note that this can be easily expanded along the first column followed by expanding the \(3\times3\) matrix which results along its first column. Thus
\[ \det(D) = 1 \begin{vmatrix} 11 & 22 \\ 14 & -17 \end{vmatrix} = 1485 \]
and so \(\det(C) = -1485\) and \(\det(A) = \det(B) = \left(\frac{-1}{3}\right)(-1485) = 495\).

**Example 4.1.29** Find the determinant of the matrix
\[ \begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & -3 & 2 & 1 \\ 2 & 1 & 2 & 5 \\ 3 & -4 & 1 & 2 \end{pmatrix} \]
Replace the second row by \((-1)\) times the first row added to it. Next take \(-2\) times the first row and add to the third and finally take \(-3\) times the first row and add to the last row. This yields
\[ \begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -1 & -1 \\ 0 & -3 & -4 & 1 \\ 0 & -10 & -8 & -4 \end{pmatrix} . \]
By Theorem 4.1.23 this matrix has the same determinant as the original matrix. Remember you can work with the columns also. Take \(-5\) times the last column and add to the second column. This yields
\[ \begin{pmatrix} 1 & -8 & 3 & 2 \\ 0 & 0 & -1 & -1 \\ 0 & -8 & -4 & 1 \\ 0 & 10 & -8 & -4 \end{pmatrix} \]
By Theorem 4.1.25 this matrix has the same determinant as the original matrix. Now take \((-1)\) times the third row and add to the top row. This gives.
\[ \begin{pmatrix} 1 & 0 & 7 & 1 \\ 0 & 0 & -1 & -1 \\ 0 & -8 & -4 & 1 \\ 0 & 10 & -8 & -4 \end{pmatrix} \]
which by Theorem 4.1.23 has the same determinant as the original matrix. Lets expand it now along the first column. This yields the following for the determinant of the original matrix.
\[ \det \begin{pmatrix} 0 & -1 & -1 \\ -8 & -4 & 1 \\ 10 & -8 & -4 \end{pmatrix} \]
which equals
\[ 8 \det \begin{pmatrix} -1 & -1 \\ -8 & -4 \end{pmatrix} + 10 \det \begin{pmatrix} -1 & -1 \\ -4 & 1 \end{pmatrix} = -82 \]
We suggest you do not try to be fancy in using row operations. That is, stick mostly to the one which replaces a row or column with a multiple of another row or column added to it. Also note there is no way to check your answer other than working the problem more than one way. To be sure you have gotten it right you must do this.

4.2 Applications

4.2.1 A Formula For The Inverse

The definition of the determinant in terms of Laplace expansion along a row or column also provides a way to give a formula for the inverse of a matrix. Recall the definition of the inverse of a matrix in Definition 3.1.28 on Page 49. Also recall the definition of the cofactor matrix given in Definition 4.1.12 on Page 60. This cofactor matrix was just the matrix which results from replacing the \( ij^{th} \) entry of the matrix with the \( ij^{th} \) cofactor.

The following theorem 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 \( A \). In other words, \( A^{-1} \) is equal to one divided by the determinant of \( A \) times the adjugate matrix of \( A \). This is what the following theorem says with more precision.

**Theorem 4.2.1** \( A^{-1} \) 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 \( \text{cof}(A)_{ij} \) the \( ij^{th} \) cofactor of \( A \).

**Example 4.2.2** Find the inverse of the matrix,

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

First find the determinant of this matrix. Using Theorems 4.1.23 - 4.1.25 on Page 62, the determinant of this matrix equals the determinant of the matrix,

\[
\begin{pmatrix} 1 & 2 & 3 \\ 0 & -6 & -8 \\ 0 & 0 & -2 \end{pmatrix}
\]

which equals 12. The cofactor matrix of \( A \) is

\[
\begin{pmatrix} -2 & -2 & 6 \\ 4 & -2 & 0 \\ 2 & 8 & -6 \end{pmatrix}
\]

Each entry of \( A \) was replaced by its cofactor. Therefore, from the above theorem, the inverse of \( A \) should equal

\[
\frac{1}{12} \begin{pmatrix} -2 & -2 & 6 \\ 4 & -2 & 0 \\ 2 & 8 & -6 \end{pmatrix}^T = \begin{pmatrix} -\frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ -\frac{1}{6} & -\frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & 0 & -\frac{1}{2} \end{pmatrix}
\]
Does it work? You should check to see if it does. When the matrices are multiplied
\[
\begin{pmatrix}
-\frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\
-\frac{1}{6} & -\frac{1}{2} & \frac{1}{3} \\
\frac{1}{2} & 0 & -\frac{1}{2}
\end{pmatrix}
\begin{pmatrix}
1 & 2 & 3 \\
3 & 0 & 1 \\
1 & 2 & 1
\end{pmatrix}
= \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]
and so it is correct.

**Example 4.2.3** Find the inverse of the matrix,

\[
A = \begin{pmatrix}
\frac{1}{2} & 0 & \frac{1}{2} \\
-\frac{1}{6} & \frac{1}{3} & -\frac{1}{2} \\
-\frac{5}{6} & \frac{2}{3} & -\frac{1}{2}
\end{pmatrix}
\]

First find its determinant. This determinant is \( \frac{1}{6} \). The inverse is therefore equal to

\[
6 \begin{pmatrix}
\frac{1}{3} & -\frac{1}{2} & -\frac{1}{6} & -\frac{1}{2} & -\frac{1}{6} & \frac{1}{3} \\
\frac{2}{3} & -\frac{1}{2} & -\frac{5}{6} & -\frac{1}{2} & -\frac{5}{6} & \frac{2}{3} \\
0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\
\frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\
-\frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\
-\frac{1}{6} & \frac{1}{3} & \frac{1}{6} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{3}
\end{pmatrix}^T.
\]

Expanding all the \( 2 \times 2 \) determinants this yields

\[
6 \begin{pmatrix}
\frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\
\frac{1}{3} & \frac{1}{6} & -\frac{1}{3} \\
-\frac{1}{6} & \frac{1}{6} & \frac{1}{6}
\end{pmatrix}^T = \begin{pmatrix}
1 & 2 & -1 \\
2 & 1 & 1 \\
1 & -2 & 1
\end{pmatrix}
\]

Always check your work.

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

and so we got it right. If the result of multiplying these matrices had been something other than the identity matrix, you would know there was an error. When this happens, you need to search for the mistake if you are interested in getting the right answer. A common mistake is to forget to take the transpose of the cofactor matrix.
Proof of Theorem 4.2.1: From the definition of the determinant in terms of expansion along a column, and letting \((a_{ir}) = A\), if \(\det(A) \neq 0\),
\[
\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ir} \det(A)^{-1} = \det(A) \det(A)^{-1} = 1.
\]

Now consider
\[
\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1}
\]
when \(k \neq r\). Replace the \(k^{th}\) column with the \(r^{th}\) column to obtain a matrix, \(B_k\) whose determinant equals zero by Theorem 4.1.21. However, expanding this matrix, \(B_k\) along the \(k^{th}\) column yields
\[
0 = \det(B_k) \det(A)^{-1} = \sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1}
\]
Summarizing,
\[
\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1} = \delta_{rk} \equiv \begin{cases} 
1 & \text{if } r = k \\
0 & \text{if } r \neq k 
\end{cases}
\]
Now
\[
\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} = \sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ki}^T
\]
which is the \(kr^{th}\) entry of \(\operatorname{cof}(A)^T A\). Therefore,
\[
\frac{\operatorname{cof}(A)^T}{\det(A)} A = I. \tag{4.2}
\]

Using the other formula in Definition 4.1.13, and similar reasoning,
\[
\sum_{j=1}^{n} a_{rj} \operatorname{cof}(A)_{kj} \det(A)^{-1} = \delta_{rk}
\]
Now
\[
\sum_{j=1}^{n} a_{rj} \operatorname{cof}(A)_{kj} = \sum_{j=1}^{n} a_{rj} \operatorname{cof}(A)_{jk}^T
\]
which is the \(rk^{th}\) entry of \(A \operatorname{cof}(A)^T\). Therefore,
\[
A \frac{\operatorname{cof}(A)^T}{\det(A)} = I, \tag{4.3}
\]
and it follows from (4.2) and (4.3) that \(A^{-1} = (a_{ij}^{-1})\), where
\[
a_{ij}^{-1} = \operatorname{cof}(A)_{ji} \det(A)^{-1}.
\]
In other words,
\[
A^{-1} = \frac{\operatorname{cof}(A)^T}{\det(A)}.
\]
Now suppose $A^{-1}$ exists. Then by Theorem 4.1.26,

$$1 = \det(I) = \det(AA^{-1}) = \det(A)\det(A^{-1})$$

so $\det(A) \neq 0$. This proves the theorem.

This way of finding inverses is especially useful in the case where it is desired to find the inverse of a matrix whose entries are functions.

**Example 4.2.4** Suppose

$$A(t) = \begin{pmatrix} e^t & 0 & 0 \\ 0 & \cos t & \sin t \\ 0 & -\sin t & \cos t \end{pmatrix}$$

Show that $A(t)^{-1}$ exists and then find it.

First note $\det(A(t)) = e^t \neq 0$ so $A(t)^{-1}$ exists. The cofactor matrix is

$$C(t) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & -e^t \sin t & e^t \cos t \end{pmatrix}$$

and so the inverse is

$$\frac{1}{e^t} \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & -e^t \sin t & e^t \cos t \end{pmatrix}^T = \begin{pmatrix} e^{-t} & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{pmatrix}.$$
4.2. APPLICATIONS

Procedure 4.2.5 Suppose $A$ is an $n \times n$ matrix and it is desired to solve the system $Ax = y$, $y = (y_1, \cdots, y_n)^T$ for $x = (x_1, \cdots, x_n)^T$. Then Cramer’s rule says

$$x_i = \frac{\det A_i}{\det A}$$

where $A_i$ is obtained from $A$ by replacing the $i^{th}$ column of $A$ with the column $(y_1, \cdots, y_n)^T$.

Example 4.2.6 Find $x, y$ if

$$
\begin{pmatrix}
1 & 2 & 1 \\
3 & 2 & 1 \\
2 & -3 & 2
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
=
\begin{pmatrix}
1 \\
2 \\
3
\end{pmatrix}.
$$

From Cramer’s rule,

$$x = \frac{\begin{vmatrix}
1 & 2 & 1 \\
2 & 2 & 1 \\
3 & -3 & 2
\end{vmatrix}}{\begin{vmatrix}
1 & 2 & 1 \\
3 & 2 & 1 \\
2 & -3 & 2
\end{vmatrix}} = \frac{1}{2}$$

Now to find $y$,

$$y = \frac{\begin{vmatrix}
1 & 1 & 1 \\
3 & 2 & 1 \\
2 & 3 & 2
\end{vmatrix}}{\begin{vmatrix}
1 & 2 & 1 \\
3 & 2 & 1 \\
2 & -3 & 2
\end{vmatrix}} = -\frac{1}{7}$$

You see the pattern. For large systems Cramer’s rule is less than useful if you want to find an answer. This is because to use it you must evaluate determinants. However, you have no practical way to evaluate determinants for large matrices other than row operations and if you are using row operations, you might just as well use them to solve the system to begin with. It will be a lot less trouble. Nevertheless, there are situations in which Cramer’s rule is useful.

Example 4.2.7 Solve for $z$ if

$$
\begin{pmatrix}
1 & 0 & 0 \\
0 & e^t \cos t & e^t \sin t \\
0 & -e^t \sin t & e^t \cos t
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
=
\begin{pmatrix}
1 \\
t \\
t^2
\end{pmatrix}.$$
You could do it by row operations but it might be easier in this case to use Cramer’s rule because the matrix of coefficients does not consist of numbers but of functions. Thus

\[
z = \begin{vmatrix}
1 & 0 & 1 \\
0 & e^t \cos t & t \\
0 & -e^t \sin t & t^2
\end{vmatrix}
+ \begin{vmatrix}
1 & 0 & 0 \\
0 & e^t \cos t & 0 \\
0 & -e^t \sin t & e^t \cos t
\end{vmatrix} = t ((\cos t) t + \sin t) e^{-t}.
\]

You end up doing this sort of thing sometimes in ordinary differential equations in the method of variation of parameters.

### 4.3 Exercises

1. Find the determinants of the following matrices.
   
   (a) \[
   \begin{pmatrix}
   1 & 2 & 3 \\
   3 & 2 & 2 \\
   0 & 9 & 8
   \end{pmatrix}
   \] (The answer is 31.)

   (b) \[
   \begin{pmatrix}
   4 & 3 & 2 \\
   1 & 7 & 8 \\
   3 & -9 & 3
   \end{pmatrix}
   \] (The answer is 375.)

   (c) \[
   \begin{pmatrix}
   1 & 2 & 3 & 2 \\
   4 & 1 & 5 & 0 \\
   1 & 2 & 1 & 2
   \end{pmatrix}
   , \) (The answer is −2.)

2. Find the following determinant by expanding along the first row and second column.

   \[
   \begin{vmatrix}
   1 & 2 & 1 \\
   2 & 1 & 3 \\
   2 & 1 & 1
   \end{vmatrix}
   \]

3. Find the following determinant by expanding along the first column and third row.

   \[
   \begin{vmatrix}
   1 & 2 & 1 \\
   1 & 0 & 1 \\
   2 & 1 & 1
   \end{vmatrix}
   \]

4. Find the following determinant by expanding along the second row and first column.

   \[
   \begin{vmatrix}
   1 & 2 & 1 \\
   2 & 1 & 3 \\
   2 & 1 & 1
   \end{vmatrix}
   \]

5. Compute the determinant by cofactor expansion. Pick the easiest row or column to use.

   \[
   \begin{vmatrix}
   1 & 0 & 0 & 1 \\
   2 & 1 & 1 & 0 \\
   0 & 0 & 0 & 2 \\
   2 & 1 & 3 & 1
   \end{vmatrix}
   \]
6. Find the determinant using row operations.
\[
\begin{vmatrix}
1 & 2 & 1 \\
2 & 3 & 2 \\
-4 & 1 & 2 \\
\end{vmatrix}
\]

7. Find the determinant using row operations.
\[
\begin{vmatrix}
2 & 1 & 3 \\
2 & 4 & 2 \\
1 & 4 & -5 \\
\end{vmatrix}
\]

8. Find the determinant using row operations.
\[
\begin{vmatrix}
1 & 2 & 1 & 2 \\
3 & 1 & -2 & 3 \\
-1 & 0 & 3 & 1 \\
2 & 3 & 2 & -2 \\
\end{vmatrix}
\]

9. Find the determinant using row operations.
\[
\begin{vmatrix}
1 & 4 & 1 & 2 \\
3 & 2 & -2 & 3 \\
-1 & 0 & 3 & 3 \\
2 & 1 & 2 & -2 \\
\end{vmatrix}
\]

10. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.
\[
\begin{pmatrix}
a & b \\
c & d \\
\end{pmatrix}, \begin{pmatrix}
a & c \\
b & d \\
\end{pmatrix}
\]

11. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.
\[
\begin{pmatrix}
a & b \\
c & d \\
\end{pmatrix}, \begin{pmatrix}
c & d \\
a & b \\
\end{pmatrix}
\]

12. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.
\[
\begin{pmatrix}
a & b \\
c & d \\
\end{pmatrix}, \begin{pmatrix}
a & b \\
a+c & b+d \\
\end{pmatrix}
\]

13. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.
\[
\begin{pmatrix}
a & b \\
c & d \\
\end{pmatrix}, \begin{pmatrix}
a & b \\
2c & 2d \\
\end{pmatrix}
\]

14. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.
\[
\begin{pmatrix}
a & b \\
c & d \\
\end{pmatrix}, \begin{pmatrix}
b & a \\
d & c \\
\end{pmatrix}
\]
DETERMINANTS

15. Tell whether the statement is true or false.

(a) If $A$ is a $3 \times 3$ matrix with a zero determinant, then one column must be a multiple of some other column.

(b) If any two columns of a square matrix are equal, then the determinant of the matrix equals zero.

(c) For $A$ and $B$ two $n \times n$ matrices, $\det(A + B) = \det(A) + \det(B)$.

(d) For $A$ an $n \times n$ matrix, $\det(3A) = 3\det(A)$

(e) If $A^{-1}$ exists then $\det(A^{-1}) = \det(A)^{-1}$.

(f) If $B$ is obtained by multiplying a single row of $A$ by 4 then $\det(B) = 4\det(A)$.

(g) For $A$ an $n \times n$ matrix, $\det(-A) = (-1)^n \det(A)$.

(h) If $A$ is a real $n \times n$ matrix, then $\det(A^T A) \geq 0$.

(i) Cramer’s rule is useful for finding solutions to systems of linear equations in which there is an infinite set of solutions.

(j) If $A^k = 0$ for some positive integer, $k$, then $\det(A) = 0$.

(k) If $Ax = 0$ for some $x \neq 0$, then $\det(A) = 0$.

16. Verify an example of each property of determinants found in Theorems 4.1.23 - 4.1.25 for $2 \times 2$ matrices.

17. A matrix is said to be orthogonal if $A^T A = I$. Thus the inverse of an orthogonal matrix is just its transpose. What are the possible values of $\det(A)$ if $A$ is an orthogonal matrix?

18. Fill in the missing entries to make the matrix orthogonal as in Problem 17.

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

19. If $A^{-1}$ exist, what is the relationship between $\det(A)$ and $\det(A^{-1})$. Explain your answer.

20. Is it true that $\det(A + B) = \det(A) + \det(B)$? If this is so, explain why it is so and if it is not so, give a counter example.

21. Let $A$ be an $r \times r$ matrix and suppose there are $r - 1$ rows (columns) such that all rows (columns) are linear combinations of these $r - 1$ rows (columns). Show $\det(A) = 0$.

22. Show $\det(aA) = a^n \det(A)$ where here $A$ is an $n \times n$ matrix and $a$ is a scalar.

23. Suppose $A$ is an upper triangular matrix. Show that $A^{-1}$ exists if and only if all elements of the main diagonal are non zero. Is it true that $A^{-1}$ will also be upper triangular? Explain. Is everything the same for lower triangular matrices?

24. Let $A$ and $B$ be two $n \times n$ matrices. $A \sim B$ (A is similar to B) means there exists an invertible matrix, $S$ such that $A = S^{-1}BS$. Show that if $A \sim B$, then $B \sim A$. Show also that $A \sim A$ and that if $A \sim B$ and $B \sim C$, then $A \sim C$. 

25. In the context of Problem 24 show that if \( A \sim B \), then \( \det(A) = \det(B) \).

26. Two \( n \times n \) matrices, \( A \) and \( B \), are similar if \( B = S^{-1}AS \) for some invertible \( n \times n \) matrix, \( S \). Show that if two matrices are similar, they have the same characteristic polynomials. The characteristic polynomial of an \( n \times n \) matrix, \( M \) is the polynomial, \( \det(M - \lambda I) \).

27. Prove by doing computations that \( \det(AB) = \det(A)\det(B) \) if \( A \) and \( B \) are \( 2 \times 2 \) matrices.

28. Illustrate with an example of \( 2 \times 2 \) matrices that the determinant of a product equals the product of the determinants.

29. An \( n \times n \) matrix is called nilpotent if for some positive integer, \( k \) it follows \( A^k = 0 \). If \( A \) is a nilpotent matrix and \( k \) is the smallest possible integer such that \( A^k = 0 \), what are the possible values of \( \det(A) \)?

30. Use Cramer’s rule to find the solution to
\[
\begin{align*}
x + 2y &= 1 \\
2x - y &= 2
\end{align*}
\]

31. Use Cramer’s rule to find the solution to
\[
\begin{align*}
x + 2y + z &= 1 \\
2x - y - z &= 2 \\
x + z &= 1
\end{align*}
\]

32. Here is a matrix,
\[
\begin{pmatrix}
1 & 2 & 3 \\
0 & 2 & 1 \\
3 & 1 & 0
\end{pmatrix}
\]
Determine whether the matrix has an inverse by finding whether the determinant is non zero.

33. Here is a matrix,
\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & \cos t & -\sin t \\
0 & \sin t & \cos t
\end{pmatrix}
\]
Does there exist a value of \( t \) for which this matrix fails to have an inverse? Explain.

34. Here is a matrix,
\[
\begin{pmatrix}
1 & t & t^2 \\
0 & 1 & 2t \\
t & 0 & 2
\end{pmatrix}
\]
Does there exist a value of \( t \) for which this matrix fails to have an inverse? Explain.

35. Here is a matrix,
\[
\begin{pmatrix}
e^t & e^{-t} \cos t & e^{-t} \sin t \\
e^t & -e^{-t} \cos t - e^{-t} \sin t & -e^{-t} \sin t + e^{-t} \cos t \\
e^t & 2e^{-t} \sin t & -2e^{-t} \cos t
\end{pmatrix}
\]
Does there exist a value of \( t \) for which this matrix fails to have an inverse? Explain.
36. Here is a matrix,
\[
\begin{pmatrix}
 e^t & \cosh t & \sinh t \\
 e^t & \sinh t & \cosh t \\
 e^t & \cosh t & \sinh t \\
\end{pmatrix}
\]
Does there exist a value of \( t \) for which this matrix fails to have an inverse? Explain.

37. Use the formula for the inverse in terms of the cofactor matrix to find if possible the inverses of the matrices
\[
\begin{pmatrix}
 1 & 1 \\
 1 & 2 \\
\end{pmatrix}, \begin{pmatrix}
 1 & 2 & 3 \\
 0 & 2 & 1 \\
 4 & 1 & 1 \\
\end{pmatrix}, \begin{pmatrix}
 1 & 2 & 1 \\
 2 & 3 & 0 \\
 0 & 1 & 2 \\
\end{pmatrix}.
\]
If it is not possible to take the inverse, explain why.

38. Use the formula for the inverse in terms of the cofactor matrix to find the inverse of the matrix,
\[
\begin{pmatrix}
 e^t & 0 & 0 \\
 0 & e^t \cos t & e^t \sin t \\
 0 & e^t \cos t - e^t \sin t & e^t \cos t + e^t \sin t \\
\end{pmatrix}.
\]

39. Find the inverse if it exists of the matrix,
\[
\begin{pmatrix}
 e^t \cos t & \sin t \\
 e^t & -\sin t & \cos t \\
 e^t & -\cos t & -\sin t \\
\end{pmatrix}.
\]

40. Let \( F(t) = \det \begin{pmatrix}
 a(t) & b(t) \\
 c(t) & d(t) \\
\end{pmatrix} \). Verify
\[
F'(t) = \det \begin{pmatrix}
 a'(t) & b'(t) \\
 c(t) & d(t) \\
\end{pmatrix} + \det \begin{pmatrix}
 a(t) & b(t) \\
 c'(t) & d'(t) \\
\end{pmatrix}.
\]
Now suppose \( F(t) = \det \begin{pmatrix}
 a(t) & b(t) & c(t) \\
 d(t) & e(t) & f(t) \\
 g(t) & h(t) & i(t) \\
\end{pmatrix} \).
Use Laplace expansion and the first part to verify \( F'(t) = \)
\[
\det \begin{pmatrix}
 a'(t) & b'(t) & c'(t) \\
 d(t) & e(t) & f(t) \\
 g(t) & h(t) & i(t) \\
\end{pmatrix} + \det \begin{pmatrix}
 a(t) & b(t) & c(t) \\
 d'(t) & e'(t) & f'(t) \\
 g(t) & h(t) & i(t) \\
\end{pmatrix} + \det \begin{pmatrix}
 a(t) & b(t) & c(t) \\
 d(t) & e(t) & f(t) \\
 g'(t) & h'(t) & i'(t) \\
\end{pmatrix}.
\]
Conjecture a general result valid for \( n \times n \) matrices and explain why it will be true.
Can a similar thing be done with the columns?

41. 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 a closed interval, \((a,b)\) and \( y \) is some function which
has $n$ derivatives so it makes sense to write $Ly$. Suppose $Ly_k = 0$ for $k = 1, 2, \cdots, n$. The Wronskian of these functions, $y_i$, is defined as

$$W(y_1, \cdots, y_n)(x) \equiv \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y'_1(x) & \cdots & y'_n(x) \\ \vdots & \cdots & \vdots \\ y_{(n-1)}(x) & \cdots & y_{(n-1)}(x) \end{pmatrix}$$

Show that for $W(x) = W(y_1, \cdots, y_n)(x)$ to save space,

$$W'(x) = \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y'_1(x) & \cdots & y'_n(x) \\ \vdots & \cdots & \vdots \\ y_{(n)}(x) & \cdots & y_{(n)}(x) \end{pmatrix}.$$  

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. **Hint:** To solve the differential equation, let $A'(x) = a_{n-1}(x)$ and multiply both sides of the differential equation by $e^{A(x)}$ and then argue the left side is the derivative of something.

### 4.4 The Mathematical Theory Of Determinants

It is easiest to give a different definition of the determinant which is clearly well defined and then prove the earlier one in terms of Laplace expansion. 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. There will be some repetition between this section and the earlier section on determinants. The main purpose is to give all the missing proofs. Two books which give a good introduction to determinants are Apostol [2] and Rudin [22]. A recent book which also has a good introduction is Baker [4].

The following Lemma will be essential in the definition of the determinant.

**Lemma 4.4.1** There exists a unique function, $\text{sgn}_n$ which maps each list of numbers from $\{1, \cdots, n\}$ to one of the three numbers, $0, 1$, or $-1$ which also has the following properties.

\[
\text{sgn}_n(1, \cdots, n) = 1 \\
\text{sgn}_n(i_1, \cdots, p, \cdots, q, \cdots, i_n) = -\text{sgn}_n(i_1, \cdots, q, \cdots, p, \cdots, i_n)
\]
In words, the second property states that if two of the numbers are switched, the value of the function is multiplied by \(-1\). Also, in the case where \(n > 1\) and \(\{i_1, \ldots, i_n\} = \{1, \ldots, n\}\) so that every number from \(\{1, \ldots, n\}\) appears in the ordered list, \((i_1, \ldots, i_n)\),

\[
\text{sgn}_n (i_1, \ldots, i_{\theta-1}, n, i_{\theta+1}, \ldots, i_n) \equiv (-1)^{n-\theta} \text{sgn}_{n-1} (i_1, \ldots, i_{\theta-1}, i_{\theta+1}, \ldots, i_n)
\]  

(4.6)

where \(n = i_\theta\) in the ordered list, \((i_1, \ldots, i_n)\).

**Proof:** To begin with, it is necessary to show the existence of such a function. This is clearly true if \(n = 1\). Define \(\text{sgn}_1 (1) \equiv 1\) and observe that it works. No switching is possible.

In the case where \(n = 2\), it is also clearly true. Let \(\text{sgn}_2 (1, 2) = 1\) and \(\text{sgn}_2 (2, 1) = 0\) while \(\text{sgn}_2 (2, 2) = \text{sgn}_2 (1, 1) = 0\) and verify it works. Assuming such a function exists for \(n\), \(\text{sgn}_{n+1}\) will be defined in terms of \(\text{sgn}_n\). If there are any repeated numbers in \((i_1, \ldots, i_{n+1})\), \(\text{sgn}_{n+1} (i_1, \ldots, i_{n+1}) \equiv 0\). If there are no repeats, then \(n + 1\) appears somewhere in the ordered list. Let \(\theta\) be the position of the number \(n + 1\) in the list. Thus, the list is of the form \((i_1, \ldots, i_{\theta-1}, n + 1, i_{\theta+1}, \ldots, i_{n+1})\). From (4.6) it must be that

\[
\text{sgn}_{n+1} (i_1, \ldots, i_{\theta-1}, n + 1, i_{\theta+1}, \ldots, i_{n+1}) \equiv (-1)^{n+1-\theta} \text{sgn}_n (i_1, \ldots, i_{\theta-1}, i_{\theta+1}, \ldots, i_{n+1})\]

It is necessary to verify this satisfies (4.4) and (4.5) with \(n\) replaced with \(n + 1\). The first of these is obviously true because

\[
\text{sgn}_{n+1} (1, \ldots, n, n + 1) \equiv (-1)^{n+1-(n+1)} \text{sgn}_n (1, \ldots, n) = 1.
\]

If there are repeated numbers in \((i_1, \ldots, i_{n+1})\), then it is obvious (4.5) holds because both sides would equal zero from the above definition. It remains to verify (4.5) in the case where there are no numbers repeated in \((i_1, \ldots, i_{n+1})\). Consider

\[
\text{sgn}_{n+1} \left( i_1, \ldots, p, \ldots, q, \ldots, i_{n+1} \right),
\]

where the \(r\) above the \(p\) indicates the number, \(p\) is in the \(r^{th}\) position and the \(s\) above the \(q\) indicates that the number, \(q\) is in the \(s^{th}\) position. Suppose first that \(r < \theta < s\). Then

\[
\text{sgn}_{n+1} \left( i_1, \ldots, p, \ldots, n + 1, \ldots, q, \ldots, i_{n+1} \right) \equiv (-1)^{n+1-\theta} \text{sgn}_n \left( i_1, \ldots, p, \ldots, s-1, q, \ldots, i_{n+1} \right)
\]

while

\[
\text{sgn}_{n+1} \left( i_1, \ldots, q, \ldots, n + 1, \ldots, p, \ldots, i_{n+1} \right) = (-1)^{n+1-\theta} \text{sgn}_n \left( i_1, \ldots, r, \ldots, s-1, p, \ldots, i_{n+1} \right)
\]

and so, by induction, a switch of \(p\) and \(q\) introduces a minus sign in the result. Similarly, if \(\theta > s\) or if \(\theta < r\) it also follows that (4.5) holds. The interesting case is when \(\theta = r\) or \(\theta = s\). Consider the case where \(\theta = r\) and note the other case is entirely similar.

\[
\text{sgn}_{n+1} \left( i_1, \ldots, n + 1, \ldots, q, \ldots, i_{n+1} \right) =
\]
Let position in (4.8). By induction, each of these switches introduces a factor of $\det \{ \}$ by appropriate numbers.

In what follows $\text{sgn}$ will often be used rather than $\text{sgn}_n$ as its domain. Define $\text{sgn}_n \left( i_1, \ldots, s^{-1}q, \ldots, i_n \right) = 0$ for all possible choices of 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. Therefore, $\text{sgn}_{n+1} \left( i_1, \ldots, n + 1, \ldots, s^{-1}q, \ldots, i_n \right) = \left( -1 \right)^{n+1-r} \text{sgn}_n \left( i_1, \ldots, s^{-1}q, \ldots, i_n \right)$.

This proves the existence of the desired function.

To see this function is unique, note that you can obtain any ordered list of distinct numbers from a sequence of switches. If there exist two functions, $f$ and $g$ both satisfying (4.4) and (4.5), you could start with $f(1, \ldots, n) = g(1, \ldots, n)$ and applying the same sequence of switches, eventually arrive at $f(i_1, \ldots, i_n) = g(i_1, \ldots, i_n)$. If any numbers are repeated, then (4.5) gives both functions are equal to zero for that ordered list. This proves the lemma.

In what follows $\text{sgn}$ will often be used rather than $\text{sgn}_n$ because the context supplies the appropriate $n$.

**Definition 4.4.2** 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 4.4.3** 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 $(r_1, \cdots, r_n)$ denote an ordered list of $n$ numbers from $\{1, \cdots, n\}$. Let $A(r_1, \cdots, r_n)$ denote the matrix whose $k^{th}$ row is the $r_k$ row of the matrix, $A$. Thus

$$\det (A(r_1, \cdots, r_n)) = \sum_{(k_1, \cdots, k_n)} \text{sgn} (k_1, \cdots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n} \quad (4.9)$$

and

$$A(1, \cdots, n) = A.$$

Proposition 4.4.4 Let $(r_1, \cdots, r_n)$ be an ordered list of numbers from $\{1, \cdots, n\}$. Then

$$\text{sgn} (r_1, \cdots, r_n) \det (A) \quad (4.10)$$

Proof: Let $(1, \cdots, n) = (1, \cdots, r, \cdots, s, \cdots, n)$ so $r < s$.

$$\det (A(1, \cdots, r, \cdots, s, \cdots, n)) = \quad (4.12)$$

and renaming the variables, calling $k_s$, $k_r$ and $k_r$, $k_s$, this equals

$$= \sum_{(k_1, \cdots, k_n)} \text{sgn} (k_1, \cdots, k_r, \cdots, k_s, \cdots, k_n) a_{1 k_1} \cdots a_{r k_r} \cdots a_{s k_s} \cdots a_{n k_n},$$

This got switched

$$= -\sum_{(k_1, \cdots, k_n)} \text{sgn} (k_1, \cdots, k_r, \cdots, k_s, \cdots, k_n) a_{1 k_1} \cdots a_{r k_r} \cdots a_{s k_s} \cdots a_{n k_n}$$

$$= -\det (A(1, \cdots, s, \cdots, r, \cdots, n)). \quad (4.13)$$

Consequently,

$$\det (A(1, \cdots, s, \cdots, r, \cdots, n)) =$$

$$-\det (A(1, \cdots, r, \cdots, s, \cdots, n)) = -\det (A)$$

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 4.4.1, this implies

$$\det (A(r_1, \cdots, r_n)) = (-1)^p \det (A) = \text{sgn} (r_1, \cdots, r_n) \det (A)$$

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 $r^{th}$ row equals the $s^{th}$ row, then the reasoning of (4.12) - (4.13) shows that $A(r_1, \cdots, r_n) = 0$ and also $\text{sgn} (r_1, \cdots, r_n) = 0$ so the formula holds in this case also.
Observation 4.4.5 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.

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 4.4.6** The following formula for $\det(A)$ is valid.

$$\det(A) = \frac{1}{n!} \sum_{(r_1, \cdots, r_n)} \sum_{(k_1, \cdots, k_n)} \text{sgn}(r_1, \cdots, r_n) \text{sgn}(k_1, \cdots, k_n) a_{r_1k_1} \cdots a_{r_nk_n}. \quad (4.14)$$

And also $\det(A^T) = \det(A)$ where $A^T$ is the transpose of $A$. (Recall that for $A^T = (a^n_j)$, $a^n_j = a^i_j$.)

**Proof:** From Proposition 4.4.4, if the $r_i$ are distinct,

$$\det(A) = \sum_{(k_1, \cdots, k_n)} \text{sgn}(r_1, \cdots, r_n) \text{sgn}(k_1, \cdots, k_n) a_{r_1k_1} \cdots a_{r_nk_n}.$$ 

Summing over all ordered lists, $(r_1, \cdots, r_n)$ where the $r_i$ are distinct, (If the $r_i$ are not distinct, $\text{sgn}(r_1, \cdots, r_n) = 0$ and so there is no contribution to the sum.)

$$n! \det(A) = \sum_{(r_1, \cdots, r_n)} \sum_{(k_1, \cdots, k_n)} \text{sgn}(r_1, \cdots, r_n) \text{sgn}(k_1, \cdots, k_n) a_{r_1k_1} \cdots a_{r_nk_n}.$$ 

This proves the corollary since the formula gives the same number for $A$ as it does for $A^T$.

**Corollary 4.4.7** 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 $i^{th}$ 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 $i^{th}$ row of $A_1$ is $(a_1, \cdots, a_n)$ and the $i^{th}$ 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”.

**Proof:** By Proposition 4.4.4 when two rows are switched, the determinant of the resulting matrix is $(-1)$ times the determinant of the original matrix. By Corollary 4.4.6 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,

$$\det(A) = \det(A^T) = -\det(A^T) = -\det(A_1).$$ 

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 $\det(A) = 0$. 

4.4. THE MATHEMATICAL THEORY OF DETERMINANTS 79
DETERMINANTS

It remains to verify the last assertion.

\[
\det(A) \equiv \sum_{(k_1, \cdots, k_n)} \text{sgn}(k_1, \cdots, k_n) a_{1k_1} \cdots (xa_{k_i} + yb_{k_i}) \cdots a_{nk_n}
\]

\[
= x \sum_{(k_1, \cdots, k_n)} \text{sgn}(k_1, \cdots, k_n) a_{1k_1} \cdots a_{k_i} \cdots a_{nk_n}
\]

\[
+ y \sum_{(k_1, \cdots, k_n)} \text{sgn}(k_1, \cdots, k_n) a_{1k_1} \cdots b_{k_i} \cdots a_{nk_n}
\]

\[
\equiv x \det(A_1) + y \det(A_2).
\]

The same is true of columns because \( \det(A^T) = \det(A) \) and the rows of \( A^T \) are the columns of \( A \).

**Definition 4.4.8** A vector, \( \mathbf{w} \), is a linear combination of the vectors \( \{ \mathbf{v}_1, \cdots, \mathbf{v}_r \} \) if there exists scalars, \( c_1, \cdots c_r \) such that \( \mathbf{w} = \sum_{k=1}^{r} c_k \mathbf{v}_k \). This is the same as saying \( \mathbf{w} \in \text{span} \{ \mathbf{v}_1, \cdots, \mathbf{v}_r \} \).

The following corollary is also of great use.

**Corollary 4.4.9** 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 \( A = (\mathbf{a}_1 \cdots \mathbf{a}_n) \) be the columns of \( A \) and suppose the condition that one column is a linear combination of \( r \) of the others is satisfied. Then by using Corollary 4.4.7 you may rearrange the columns to have the \( n^{th} \) column a linear combination of the first \( r \) columns. Thus \( \mathbf{a}_n = \sum_{k=1}^{r} c_k \mathbf{a}_k \) and so

\[
\det(A) = \det \left( \begin{array}{cccc}
\mathbf{a}_1 & \cdots & \mathbf{a}_r & \cdots & \mathbf{a}_{n-1} & \sum_{k=1}^{r} c_k \mathbf{a}_k
\end{array} \right).
\]

By Corollary 4.4.7

\[
\det(A) = \sum_{k=1}^{r} c_k \det \left( \begin{array}{cccc}
\mathbf{a}_1 & \cdots & \mathbf{a}_r & \cdots & \mathbf{a}_{n-1} & \mathbf{a}_k
\end{array} \right) = 0.
\]

The case for rows follows from the fact that \( \det(A) = \det(A^T) \). This proves the corollary.

Recall the following definition of matrix multiplication.

**Definition 4.4.10** If \( A \) and \( B \) are \( n \times n \) matrices, \( A = (a_{ij}) \) and \( B = (b_{ij}) \), \( AB = (c_{ij}) \) where

\[
c_{ij} \equiv \sum_{k=1}^{n} a_{ik} b_{kj}.
\]

One of the most important rules about determinants is that the determinant of a product equals the product of the determinants.

**Theorem 4.4.11** Let \( A \) and \( B \) be \( n \times n \) matrices. Then

\[
\det(AB) = \det(A) \det(B).
\]
4.4. THE MATHEMATICAL THEORY OF DETERMINANTS

**Proof:** Let $c_{ij}$ be the $ij^{th}$ entry of $AB$. Then by Proposition 4.4.4,

$$
\det (AB) = \sum_{(k_1, \ldots, k_n)} \text{sgn} (k_1, \ldots, k_n) c_{i_1k_1} \cdots c_{i_nk_n}
$$

$$
= \sum_{(k_1, \ldots, k_n)} \text{sgn} (k_1, \ldots, k_n) \left( \sum_{r_1} a_{r_11} b_{r_1k_1} \right) \cdots \left( \sum_{r_n} a_{nr_n} b_{r_nk_n} \right)
$$

$$
= \sum_{(r_1 \cdots r_n)} \sum_{(k_1, \ldots, k_n)} \text{sgn} (k_1, \ldots, k_n) b_{r_1k_1} \cdots b_{r_nk_n} \left( a_{r_11} \cdots a_{nr_n} \right)
$$

$$
= \sum_{(r_1 \cdots r_n)} \text{sgn} (r_1, \ldots, r_n) a_{r_11} \cdots a_{nr_n} \det (B) = \det (A) \det (B).
$$

This proves the theorem.

**Lemma 4.4.12** Suppose a matrix is of the form

$$
M = \begin{pmatrix} A & * \\ 0 & a \end{pmatrix}
$$

(4.15)

or

$$
M = \begin{pmatrix} A & 0 \\ * & a \end{pmatrix}
$$

(4.16)

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

$$
\det (M) = a \det (A).
$$

**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, $m_{nn} = a$ and $m_{in} = 0$ if $i \neq n$. From the definition of the determinant,

$$
\det (M) \equiv \sum_{(k_1, \ldots, k_n)} \text{sgn}_n (k_1, \ldots, k_n) m_{1k_1} \cdots m_{nk_n}
$$

Letting $\theta$ denote the position of $n$ in the ordered list, $(k_1, \ldots, k_n)$ then using the earlier conventions used to prove Lemma 4.4.1, $\det (M)$ equals

$$
\sum_{(k_1, \ldots, k_n)} (-1)^{n-\theta} \text{sgn}_{n-1} (k_1, \ldots, k_{\theta-1}, k_{\theta+1}, \ldots, k_{n-1}) m_{1k_1} \cdots m_{nk_n}
$$

Now suppose (4.16). 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

$$
a \sum_{(k_1, \ldots, k_{n-1})} \text{sgn}_{n-1} (k_1, \ldots, k_{n-1}) m_{1k_1} \cdots m_{(n-1)k_{n-1}} = a \det (A).
$$

To get the assertion in the situation of (4.15) use Corollary 4.4.6 and (4.16) to write

$$
\det (M) = \det (M^T) = \det \left( \begin{pmatrix} A^T & 0 \\ * & a \end{pmatrix} \right) = a \det (A^T) = a \det (A).
$$
This proves the lemma.

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 4.4.13** 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.

The following is the main result. Earlier this was given as a definition and the outrageous totally unjustified assertion was made that the same number would be obtained by expanding the determinant along any row or column. The following theorem proves this assertion.

**Theorem 4.4.14** Let \( A \) be an \( n \times n \) matrix where \( n \geq 2 \). Then

\[
\det(A) = \sum_{j=1}^{n} a_{ij} \text{cof}(A)_{ij} = \sum_{i=1}^{n} a_{ij} \text{cof}(A)_{ij}.
\]

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 \((a_{11}, \cdots, a_{nn})\) 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 \((0, \cdots, 0, a_{ij}, 0, \cdots, 0)\). Then by Corollary 4.4.7,

\[
\det(A) = \sum_{j=1}^{n} \det(B_j)
\]

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 \( \text{cof}(A)_{ij} \equiv (-1)^{i+j} \det(A^{ij}) \). At this point, recall that from Proposition 4.4.4, 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 4.4.12,

\[
\det(B_j) = (-1)^{n-j} (-1)^{n-i} \det\left( \begin{pmatrix} A^{ij} & * \\ 0 & a_{ij} \end{pmatrix} \right) = (-1)^{i+j} \det\left( \begin{pmatrix} A^{ij} & * \\ 0 & a_{ij} \end{pmatrix} \right) = a_{ij} \text{cof}(A)_{ij}.
\]

Therefore,

\[
\det(A) = \sum_{j=1}^{n} a_{ij} \text{cof}(A)_{ij}
\]

which is the formula for expanding \( \det(A) \) along the \( i^{th} \) row. Also,

\[
\det(A) = \det(A^T) = \sum_{j=1}^{n} a_{ij}^T \text{cof}(A^T)_{ij} = \sum_{j=1}^{n} a_{ji} \text{cof}(A)_{ji}
\]

which is the formula for expanding \( \det(A) \) along the \( i^{th} \) column. This proves the theorem.

Note that this gives an easy way to write a formula for the inverse of an \( n \times n \) matrix. Recall the definition of the inverse of a matrix in Definition 3.1.28 on Page 49.
Theorem 4.4.15 $A^{-1}$ exists if and only if $\det(A) \neq 0$. If $\det(A) \neq 0$, then $A^{-1} = (a^{-1}_{ij})$ where

$$a^{-1}_{ij} = \det(A)^{-1} \text{cof}(A)_{ji}$$

for $\text{cof}(A)_{ij}$ the $ij^{th}$ cofactor of $A$.

Proof: By Theorem 4.4.14 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 consider

$$\sum_{i=1}^{n} a_{ir} \text{cof}(A)_{ik} \det(A)^{-1}$$

when $k \neq r$. Replace the $k^{th}$ column with the $r^{th}$ column to obtain a matrix, $B_k$ whose determinant equals zero by Corollary 4.4.7. However, expanding this matrix along the $k^{th}$ column yields

$$0 = \det(B_k) \det(A)^{-1} = \sum_{i=1}^{n} a_{ir} \text{cof}(A)_{ik} \det(A)^{-1}$$

Summarizing,

$$\sum_{i=1}^{n} a_{ir} \text{cof}(A)_{ik} \det(A)^{-1} = \delta_{rk}.$$

Using the other formula in Theorem 4.4.14, 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 $A^{-1} = (a^{-1}_{ij})$, where

$$a^{-1}_{ij} = \text{cof}(A)_{ji} \det(A)^{-1}.$$

Now suppose $A^{-1}$ exists. Then by Theorem 4.4.11,

$$1 = \det(I) = \det(AA^{-1}) = \det(A) \det(A^{-1})$$

so $\det(A) \neq 0$. This proves the theorem.

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 4.4.16 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 4.4.11 implies

$$\det B \det A = 1$$

and so $\det A \neq 0$. Therefore from Theorem 4.4.15, $A^{-1}$ exists. Therefore,

$$A^{-1} = (BA) A^{-1} = B (AA^{-1}) = BI = B.$$
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 4.4.15 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,

$$x_i = \frac{1}{\det(A)} \text{det} \begin{pmatrix} * & \cdots & y_1 & \cdots & * \\ \vdots & & \vdots & & \vdots \\ * & \cdots & y_n & \cdots & * \end{pmatrix},$$

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 4.4.17** 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.

$$\begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & \vdots \\ \vdots & \cdots & \cdots & * \\ 0 & \cdots & 0 & * \end{pmatrix}$$

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 4.4.14.

**Corollary 4.4.18** Let $M$ be an upper (lower) triangular matrix. Then $\det(M)$ is obtained by taking the product of the entries on the main diagonal.

**Definition 4.4.19** 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 4.4.20** If $A$ 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.
4.4. The mathematical theory of determinants

Proof: Suppose the determinant rank of \( A = (a_{ij}) \) equals \( r \). If rows and columns are interchanged, the determinant rank of the modified matrix is unchanged. Thus rows and columns can be interchanged to produce an \( r \times r \) matrix in the upper left corner of the matrix which has non zero determinant. Now consider the \( r + 1 \times r + 1 \) matrix, \( M \),

\[
\begin{pmatrix}
  a_{11} & \cdots & a_{1r} & a_{1p} \\
  \vdots & \ddots & \vdots & \vdots \\
  a_{r1} & \cdots & a_{rr} & a_{rp} \\
  a_{11} & \cdots & a_{1r} & a_{1p}
\end{pmatrix}
\]

where \( C \) will denote the \( r \times r \) matrix in the upper left corner which has non zero determinant.

I claim \( \det(M) = 0 \).

There are two cases to consider in verifying this claim. First, suppose \( p > r \). Then the claim follows from the assumption that \( A \) has determinant rank \( r \). On the other hand, if \( p < r \), then the determinant is zero because there are two identical columns. Expand the determinant along the last column and divide by \( \det(C) \) to obtain

\[
a_{lp} = -\sum_{i=1}^{r} \frac{\text{cof}(M)_{ip}}{\det(C)} a_{ip}.
\]

Now note that \( \text{cof}(M)_{ip} \) does not depend on \( p \). Therefore the above sum is of the form

\[
a_{lp} = \sum_{i=1}^{r} m_i a_{ip}
\]

which shows the \( l^{th} \) row is a linear combination of the first \( r \) rows of \( A \). Since \( l \) is arbitrary, this proves the theorem.

Corollary 4.4.21 The determinant rank equals the row rank.

Proof: From Theorem 4.4.20, the row rank is no larger than the determinant rank. Could the row rank be smaller than the determinant rank? If so, there exist \( p \) rows for \( p < r \) such that the span of these \( p \) rows equals the row space. But this implies that the \( r \times r \) submatrix whose determinant is nonzero also has row rank no larger than \( p \) which is impossible if its determinant is to be nonzero because at least one row is a linear combination of the others.

Corollary 4.4.22 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 4.4.21, 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 4.4.23 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 4.4.20. In particular, it follows that for some m, the mth column is a linear combination of all the others. Thus letting $A = (a_1 \cdots a_m \cdots a_n)$ 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, $x \equiv (\alpha_1 \cdots -1 \cdots \alpha_n)^T$. 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

$$A^Tx = 0.$$ 

Taking the transpose of both sides yields

$$x^TA = 0$$

where the 0 is a $1 \times n$ matrix or row vector. Now if $Ay = x$, then

$$|x|^2 = x^T (Ay) = (x^TA)y = 0y = 0$$

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 4.4.15 $A^{-1}$ exists and so for every $y \in F^n$ there exists a unique $x \in F^n$ such that $Ax = y$. In fact $x = A^{-1}y$. Thus A would be onto contrary to 3). This shows 3) implies 1) and proves the theorem.

Corollary 4.4.24 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.

4.5 The Cayley Hamilton Theorem

Definition 4.5.1 Let $A$ be an $n \times n$ matrix. The characteristic polynomial is defined as

$$p_A(t) \equiv \det(tI - A).$$

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

$$p(A) \equiv A^n + a_{n-1}A^{n-1} + \cdots + a_1 A + a_0I.$$ 

The explanation for the last term is that $A^0$ is interpreted as I, the identity matrix.
4.5. THE CAYLEY HAMILTON THEOREM

The Cayley Hamilton theorem states that every matrix satisfies its characteristic equation, that equation defined by \( P_A(t) = 0 \). It is one of the most important theorems in linear algebra. The following lemma will help with its proof.

**Lemma 4.5.2** 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:** Multiply by \( \lambda^{-m} \) to obtain

\[
A_0 \lambda^{-m} + A_1 \lambda^{-m+1} + \cdots + A_{m-1} \lambda^{-1} + A_m = 0.
\]

Now let \(|\lambda| \to \infty\) to obtain \( A_m = 0 \). With this, multiply by \( \lambda \) to obtain

\[
A_0 \lambda^{-m+1} + A_1 \lambda^{-m+2} + \cdots + A_{m-1} = 0.
\]

Now let \(|\lambda| \to \infty\) to obtain \( A_{m-1} = 0 \). Continue multiplying by \( \lambda \) and letting \( \lambda \to \infty \) to obtain that all the \( A_i = 0 \). This proves the lemma.

With the lemma, here is a simple corollary.

**Corollary 4.5.3** Let \( A_i \) and \( B_i \) be \( n \times n \) matrices and suppose

\[
A_0 + A_1 \lambda + \cdots + A_m \lambda^m = B_0 + B_1 \lambda + \cdots + B_m \lambda^m
\]

for all \(|\lambda|\) large enough. Then \( A_i = B_i \) for all \( i \). Consequently if \( \lambda \) is replaced by any \( n \times n \) matrix, the two sides will be equal. That is, for \( C \) any \( n \times n \) matrix,

\[
A_0 + A_1 C + \cdots + A_m C^m = B_0 + B_1 C + \cdots + B_m C^m.
\]

**Proof:** Subtract and use the result of the lemma.

With this preparation, here is a relatively easy proof of the Cayley Hamilton theorem.

**Theorem 4.5.4** Let \( A \) be an \( n \times n \) matrix and let \( p(\lambda) \equiv \det(\lambda I - A) \) be the characteristic polynomial. Then \( p(A) = 0 \).

**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 4.4.15

\[
C(\lambda) = p(\lambda)(\lambda I - A)^{-1}.
\]

Note that each entry in \( C(\lambda) \) is a polynomial in \( \lambda \) having degree no more than \( n - 1 \). Therefore, collecting the terms,

\[
C(\lambda) = C_0 + C_1 \lambda + \cdots + C_{n-1} \lambda^{n-1}
\]

for \( C_j \) some \( n \times n \) matrix. It follows that for all \(|\lambda|\) large enough,

\[
(A - \lambda I) \left( C_0 + C_1 \lambda + \cdots + C_{n-1} \lambda^{n-1} \right) = p(\lambda) I
\]

and so Corollary 4.5.3 may be used. It follows the matrix coefficients corresponding to equal powers of \( \lambda \) are equal on both sides of this equation. Therefore, if \( \lambda \) is replaced with \( A \), the two sides will be equal. Thus

\[
0 = (A - A) \left( C_0 + C_1 A + \cdots + C_{n-1} A^{n-1} \right) = p(A) I = p(A).
\]

This proves the Cayley Hamilton theorem.
4.6 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. Show that matrix multiplication is associative. That is, \((AB)C = A(BC)\).

3. Show the inverse of a matrix, if it exists, is unique. Thus if \( AB = BA = I \), then \( B = A^{-1} \).

4. In the proof of Theorem 4.4.15 it was claimed that \( \det (I) = 1 \). Here \( I = (\delta_{ij}) \). Prove this assertion. Also prove Corollary 4.4.18.

5. Let \( v_1, \ldots, v_n \) be vectors in \( F^n \) and let \( M(v_1, \ldots, v_n) \) denote the matrix whose \( i^{th} \) column equals \( v_i \). Define

\[
d(v_1, \ldots, v_n) \equiv \det(M(v_1, \ldots, v_n)).
\]

Prove that \( d \) is linear in each variable, (multilinear), that

\[
d(v_1, \ldots, v_i, \ldots, v_j, \ldots, v_n) = -d(v_1, \ldots, v_j, \ldots, v_i, \ldots, v_n), \quad (4.17)
\]

and

\[
d(e_1, \ldots, e_n) = 1 \quad (4.18)
\]

where here \( e_j \) is the vector in \( F^n \) which has a zero in every position except the \( j^{th} \) position in which it has a one.

6. Suppose \( f : F^n \times \cdots \times F^n \to F \) satisfies (4.17) and (4.18) and is linear in each variable. Show that \( f = d \).

7. 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.

8. If \( A = (a_{ij}) \), show \( \det(A) = \sum_{(k_1, \ldots, k_n)} \text{sgn} (k_1, \ldots, k_n) a_{k_11} \cdots a_{k_nn} \).

9. Use the result of Problem 7 to evaluate by hand the determinant

\[
\det \begin{pmatrix} 1 & 2 & 3 & 2 \\ -6 & 3 & 2 & 3 \\ 5 & 2 & 2 & 3 \\ 3 & 4 & 6 & 4 \end{pmatrix}.
\]
10. Find the inverse if it exists of the matrix,
\[
\begin{pmatrix}
e^t & \cos t & \sin t \\
e^t & -\sin t & \cos t \\
e^t & -\cos t & -\sin t \\
\end{pmatrix}
\]

11. 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 a closed interval, \((a, b)\) and \( y \) is some function which has \( n \) derivatives so it makes sense to write \( Ly \). Suppose \( Ly_k = 0 \) for \( k = 1, 2, \cdots, n \).

The Wronskian of these functions, \( y_i \), is defined as
\[
W(y_1, \cdots, y_n)(x) \equiv \det \begin{pmatrix}
y_1(x) & \cdots & y_n(x) \\
y'_1(x) & \cdots & y'_n(x) \\
\vdots & \ddots & \vdots \\
y_{(n-1)}(x) & \cdots & y_{(n-1)}(x) \\
\end{pmatrix}
\]

Show that for \( W(x) = W(y_1, \cdots, y_n)(x) \) to save space,
\[
W'(x) = \det \begin{pmatrix}
y_1(x) & \cdots & y_n(x) \\
y'_1(x) & \cdots & y'_n(x) \\
\vdots & \ddots & \vdots \\
y_{(n)}(x) & \cdots & y_{(n)}(x) \\
\end{pmatrix}
\]

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.

12. Two \( n \times n \) matrices, \( A \) and \( B \), are similar if \( B = S^{-1}AS \) for some invertible \( n \times n \) matrix, \( S \). Show that if two matrices are similar, they have the same characteristic polynomials.

13. Suppose the characteristic polynomial of an \( n \times n \) matrix, \( A \), is of the form
\[
t^n + a_{n-1}t^{n-1} + \cdots + a_1t + a_0
\]
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 \( a_0 \neq 0 \).

14. 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.
Vector Spaces

5.0.1 Outcomes

1. Define vector space.

2. Define the span of a set of vectors. Recall that a span of vectors in a vector space is a subspace.

3. Determine whether a set of vectors is a subspace.

4. Define linear independence.

5. Determine whether a set of vectors is linearly independent or linearly dependent.

6. Determine a basis and the dimension of a vector space.

5.1 Vector Spaces

The symbol, $\mathbb{R}^n$, denotes the set of $n \times 1$ matrices which have all real entries. Recall that these are also called column vectors or just vectors for short. The symbol, $\mathbb{C}^n$, denotes the set of $n \times 1$ matrices which have complex entries. Thus an example of something in $\mathbb{C}^3$ is

$$\begin{pmatrix} 1 + i \\ 2 \\ 3 - 2i \end{pmatrix}.$$  

Since every real number may be considered a complex number, it follows that every vector in $\mathbb{R}^n$ is a vector in $\mathbb{C}^n$. You will have to use $\mathbb{C}^n$ when you study differential equations. These two examples will be sufficient for many applications but not for all. It turns out that all algebraic considerations are the same for either $\mathbb{R}^n$ or $\mathbb{C}^n$ and so to avoid fussing with unenlightening details, we will denote by $\mathbb{F}^n$ the set of $n \times 1$ matrices. Here $\mathbb{F}$ will stand for either $\mathbb{R}$ or $\mathbb{C}$.

5.1.1 Vector Space Axioms

The concept of a vector space turns out to be a significant unifying idea in many different subjects. For example, $\mathbb{F}^n$, the $n \times 1$ matrices (column vectors) will turn out to be a vector space. So is the collection of all $m \times n$ matrices. In differential equations, you will see another example of a vector space consisting of certain collections of functions. Solutions to linear systems in any context turn out to involve vector spaces.
Definition 5.1.1 A vector space, $V$ is a nonempty set of objects on which are defined two operations, addition and multiplication by scalars, real or complex numbers\(^1\), satisfying the properties listed below.

1. Closure under addition.
   If $u, v \in V$, then $u + v \in V$.

2. Closure under scalar multiplication.
   If $\alpha$ is a scalar and $u \in V$, then $\alpha u \in V$.

3. The following eight vector space axioms.
   - **Commutative Law Of Addition.** For $a, b \in V$
     \[
     a + b = b + a, \tag{5.1}
     \]
   - **Associative Law for Addition.** For $a, b, c \in V$,
     \[
     (a + b) + c = a + (b + c), \tag{5.2}
     \]
   - **Existence of an Additive Identity.** There exists $0 \in V$ such that for all $a \in V$,
     \[
     a + 0 = a, \tag{5.3}
     \]
   - **Existence of an Additive Inverse.** For each $a \in V$, there exists $-a \in V$ such that
     \[
     a + (-a) = 0. \tag{5.4}
     \]
   - **Distributive law over Vector Addition.** For any scalar, $\alpha$ and any two vectors, $a, b \in V$,
     \[
     \alpha (a + b) = \alpha a + \alpha b, \tag{5.5}
     \]
   - **Distributive law over Scalar Addition.** For any two scalars, $\alpha, \beta$ and any vector $a \in V$,
     \[
     (\alpha + \beta) a = \alpha a + \beta a, \tag{5.6}
     \]
   - **Associative law for Scalar Multiplication.** For any two scalars, $\alpha, \beta$ and any vector $a \in V$,
     \[
     \alpha (\beta a) = \alpha \beta (a), \tag{5.7}
     \]
   - **Rule for Multiplication by 1.** For any vector $a \in V$,
     \[
     1a = a. \tag{5.8}
     \]

Theorem 5.1.2 Let $X$ be any vector space.

1. Then the additive identity is unique.

2. For each $u \in X$, the additive inverse $-u$ is unique.

\(^1\)For this book, the field of scalars will always be either $\mathbb{R}$ or $\mathbb{C}$ but there are many other fields used. Fields are simply algebraic objects which have the same algebraic properties as the real or complex numbers.
5.1. VECTOR SPACES

3. If \( u \in X \), then \( 0u = 0 \). That is, the scalar 0 times the vector \( u \) gives the vector \( 0 \).

4. For any \( u \in X \), \( -u = (-1)u \).

**Proof:** To see the additive identity is unique, suppose \( 0' \) is another one. Then from the properties satisfied by the additive identities,

\[
0' = 0' + 0 = 0.
\]

This verifies the first claim.

Suppose \( v \) is another additive inverse for \( u \). Then

\[
-u + u = 0, \quad v + u = 0
\]

and so

\[
-u + u = v + u.
\]

Add \(-u\) to both sides as follows

\[
(-u + u) + (-u) = (v + u) + (-u).
\]

By the associative law of addition,

\[
-u + (u + (-u)) = v + (u + (-u))
\]

and so

\[
-u = -u + 0 = v + 0 = v.
\]

This verifies the second assertion.

By the distributive law,

\[
0u = (0 + 0)u = 0u + 0u.
\]

Now by the existence of the additive inverse, there exists an additive inverse to \( 0u \). Add it to both sides and then use the associative law for addition to write

\[
0 = (-0u) + 0u = (-0u) + (0u + 0u) = (-0u) + 0u + 0u = 0 + 0u = 0u.
\]

This verifies the third assertion.

To verify the fourth, use the distributive law and the third assertion which was just proved to write

\[
u + (-1)u = (1 + (-1))u = 0u = 0.
\]

Therefore, \((-1)u\) acts like the additive inverse of \( u \). By the second assertion which was just proved, it follows \((-1)u = -u\).

What follows is a list of examples. You will benefit greatly from verifying the claim that each is a vector space.

**Example 5.1.3** \( \mathbb{F}^n \) is a vector space with respect to the scalar multiplication defined earlier. Recall the properties of matrix addition and multiplication by scalars.

**Example 5.1.4** The set of \( m \times n \) matrices is a vector space.
Example 5.1.5 Consider the space of real valued functions defined on some set, \( D \). If \( f, g \) are two such functions then \( f + g \) is the function defined by \( (f + g)(x) = f(x) + g(x) \). Also for \( \alpha \) a real number, \( \alpha f \) is the function defined by \( \alpha f(x) = \alpha (f(x)) \). With this understanding of addition and scalar multiplication, this set of functions is a vector space, the field of scalars being \( \mathbb{R} \).

Example 5.1.6 Consider the space of complex valued functions defined on some set, \( D \). If \( f, g \) are two such functions then \( f + g \) is the function defined by \( (f + g)(x) = f(x) + g(x) \). Also for \( \alpha \) a complex number, \( \alpha f \) is the function defined by \( \alpha f(x) = \alpha (f(x)) \). With this understanding of addition and scalar multiplication, this set of functions is a vector space. Note that in this case the field of scalars is \( \mathbb{C} \).

Example 5.1.7 Consider the space of functions defined on \( \mathbb{R} \) which have a continuous derivative. This is a vector space.

Example 5.1.8 Consider the space of functions defined on \([0, 1]\) which have a Riemann integral. This is a vector space.

Example 5.1.9 The space of functions, \( f \) defined on an interval, \([a, b]\) which satisfy a condition of the form

\[
|f(x) - f(y)| \leq C |x - y|^\alpha
\]

for \( \alpha \in (0, 1) \) is a very important example of a vector space of functions.

Example 5.1.10 Consider the polynomials defined on \([0, 1]\) having degree no larger than 3. If you add two such functions, you get another such function and if you multiply one by a scalar, you get another one also. Also all the vector space axioms hold for this set of functions so this is also a vector space.

Example 5.1.11 Consider functions of the form \( a \sin x + b \cos x \) where \( a, b \) are real numbers. This is a vector space also. Things like this will be important in beginning courses in differential equations.

Example 5.1.12 This example is a little more exotic than the above and we won’t need it in what follows but if you are interested, let your field of scalars be the rational numbers and let your vectors be numbers of the form \( a + b\sqrt{5} \) where \( a, b \) are rational numbers. You can verify this is also a vector space.

5.1.2 Spans

An important concept in the applications of vector spaces is that of linear combinations and spans.

Definition 5.1.13 Let \( \{x_1, \ldots, x_p\} \) be vectors in a vector space, \( X \). A linear combination is any expression of the form

\[
c_1x_1 + \cdots + c_px_p = \sum_{i=1}^{p} c_ix_i
\]

where the \( c_i \) are scalars.
Example 5.1.14  The vector, \( \begin{pmatrix} 7 \\ 4 \end{pmatrix} \) is a linear combination of the vectors \( \begin{pmatrix} 1 \\ 2 \end{pmatrix} \) and \( \begin{pmatrix} 5 \\ 0 \end{pmatrix} \) because
\[
2 \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 5 \\ 0 \end{pmatrix} = \begin{pmatrix} 7 \\ 4 \end{pmatrix}.
\]

Definition 5.1.15  The set of all linear combinations of \( \{x_1, \ldots, x_p\} \) is called the span and is written as
\[
\text{span}(x_1, \ldots, x_n).
\]
You can consider the span of any number of vectors.

Example 5.1.16  Consider the span of one vector in \( \mathbb{R}^3 \).

You see there is a vector, \( \mathbf{v} \) and the span of this single vector, \( \{t\mathbf{v} \text{ such that } t \in \mathbb{R}\} \) gives the indicated line which goes through the origin, \((0,0,0)\) having \( \mathbf{v} \) as a direction vector.

Example 5.1.17  You can get an idea of the appearance of the span of two vectors in \( \mathbb{R}^3 \). These are just planes which pass through the origin. Here is a picture.

Lets consider why the displayed plane really is the span of the two vectors which lie in this plane as shown.
As indicated in the above picture, a typical thing in the span of these two vectors is of the form \( s \mathbf{u} + t \mathbf{v} \) where \( s \) and \( t \) are real numbers. By specifying \( s \), you determine a point on the line through the origin, \((0,0,0)\) having direction vector, \( \mathbf{u} \). Then through this point, there is a line having direction vector, \( \mathbf{v} \). We have drawn three such lines in the above picture, one for \( s = 0 \), \( s_1 \), and \( s_2 \). The totality of all such lines yields the span of the two vectors, \( \mathbf{u} \) and \( \mathbf{v} \) and you see from geometric considerations it is just a plane.

It is important that neither of the two vectors \( \mathbf{u} \) and \( \mathbf{v} \) be a multiple of the other in order for a plane like the one shown to be obtained. If \( \mathbf{v} \) had pointed in the same direction or opposite direction as \( \mathbf{u} \), you see that the span of these two vectors would reduce to nothing more than the line \( s \mathbf{u} \) where \( s \in \mathbb{R} \). Also if one of these is a multiple of the other, then the vector, \( \mathbf{n} \) given above would equal \( \mathbf{0} \) and so the set of vectors perpendicular to this vector \( \mathbf{n} \) would not equal a plane.

Some examples have absolutely nothing to do with geometry.

**Example 5.1.18** Let \( X \) denote the set of functions defined on an interval, \((a,b)\). Let \( f_0 (x) = 1, f_1 (x) = x, \) and \( f_2 (x) = x^2 \). Find \( \text{span} \ (f_0, f_1, f_2) \).

Something in the span is of the form \( a f_0 + b f_1 + c f_2 \) and \( (a f_0 + b f_1 + c f_2) (x) = a+bx+cx^2 \) so you see that the span of these vectors consists polynomials of degree no more than 2.
5.1. SUBSPACES

5.1.3 Subspaces

Any time you are dealing with a vector space, there is a concept of **subspace**.

**Definition 5.1.19** Let \( X \) be a vector space and let \( V \) be a collection of vectors of \( X \). Then \( V \) is called a **subspace** of \( X \) if \( V \) is a vector space contained in \( X \) with respect to the same vector addition and scalar multiplication.

**Proposition 5.1.20** Let \( X \) be a vector space and let \( V \) be a collection of vectors of \( X \). Then \( V \) is a subspace of \( X \) if and only if whenever \( \alpha, \beta \) are scalars and \( u, v \) are vectors of \( V \), it follows \( \alpha u + \beta v \in V \). That is, \( V \) is “closed under the algebraic operations of vector addition and scalar multiplication”.

**Proof:** Suppose first that \( V \) is a subspace of \( X \). This means that \( V \) is itself a vector space. Therefore, if \( \alpha, \beta \) are scalars and \( u, v \) are vectors in \( V \) it follows that \( \alpha u + \beta v \) is a vector of \( V \) from the axioms of a vector space given in Definition 5.1.1 on Page 92.

Next suppose the condition involving \( V \) being closed with respect to the vector space operations holds. Then it follows \( V \) must be a vector space because the other eight vector space axioms all are valid because they are valid for all vectors from \( X \) and so in particular they hold for vectors in \( V \) because every vector in \( V \) is given to be in \( X \). Therefore, \( V \) is a vector space. This proves the proposition.

**Example 5.1.21** For example, in \( \mathbb{R}^3 \), the subspaces are the zero vector, \( \{0\} \), any line through the origin, or any plane through the origin or all of \( \mathbb{R}^3 \). It follows from some of the theorems presented below that these are the only examples of subspaces in \( \mathbb{R}^3 \).

**Proposition 5.1.22** Let \( \{x_1, \ldots, x_p\} \) be vectors in a vector space, \( X \). Then \( \text{span} (x_1, \ldots, x_p) \) is a subspace.

**Proof:** It is suffices to verify the condition of closure with respect to the vector space operations found in Proposition 5.1.20. Let \( u, v \in \text{span} (x_1, \ldots, x_p) \). This means there exist scalars, \( \alpha_1, \ldots, \alpha_p \) and \( \beta_1, \ldots, \beta_p \) such that

\[
u = \alpha_1 x_1 + \cdots + \alpha_p x_p \equiv \sum_{i=1}^{p} \alpha_i x_i, \quad v = \sum_{i=1}^{p} \beta_i x_i.\]

If \( a, b \) are two scalars, we need to verify that \( au + bv \in \text{span} (x_1, \ldots, x_p) \). But

\[
au + bv = a \sum_{i=1}^{p} \alpha_i x_i + b \sum_{i=1}^{p} \beta_i x_i
= \sum_{i=1}^{p} (a\alpha_i + b\beta_i) x_i 
\in \text{span} (x_1, \ldots, x_p).
\]

This verifies the assertion of this proposition.

You should note that there are only finitely many vectors in \( \{x_1, \ldots, x_p\} \) but there will likely be infinitely many vectors in \( \text{span} (x_1, \ldots, x_p) \) so don’t confuse the two.

When \( S \) is any subset of a vector space, \( X \), the term \( \text{span} (S) \) denotes the set of finite linear combinations of vectors of \( S \). Suppose \( V \) is a subspace of \( X \). Then \( \text{span} (V) = V \) from the conclusion of Proposition 5.1.20. Therefore, every subspace is the span of some set of vectors. However, it is much more desirable to find a minimal set of vectors, hopefully a finite set, which also spans the subspace. We will address this issue in Section 5.1.5.
Definition 5.1.23 When \( V = \text{span} (x_1, \ldots, x_p) \), the set of vectors, \( \{x_1, \ldots, x_p\} \) is called a \textit{spanning set} for \( V \).

Example 5.1.24 Let \( X = \mathbb{F}^n \) and let \( V = \{ x = (x_1, \ldots, x_n) \in \mathbb{F}^n : x_n = 0 \} \). Is \( V \) a subspace?

You have to verify that if \( a \) and \( b \) are scalars and \( x, y \) vectors in \( V \), then the linear combination \( ax + by \) is in \( V \). Since \( x, y \) are in \( V \), it follows \( x = (x_1, \ldots, x_{n-1}, 0) \) and \( y = (y_1, \ldots, y_{n-1}, 0) \). Therefore,

\[
ax + by = a(x_1, \ldots, x_{n-1}, 0) + b(y_1, \ldots, y_{n-1}, 0)
= (ax_1 + by_1, \ldots, ax_{n-1} + by_{n-1}, 0)
\]

which is a vector of \( V \). Therefore, \( V \) is a subspace.

Example 5.1.25 Let \( X = \mathbb{F}^n \) and let \( V = \{ x = (x_1, \ldots, x_n) \in \mathbb{F}^n : x_n \geq 0 \} \). Is \( V \) a subspace?

In this case, \( V \) is not a subspace because \( (x_1, \ldots, x_{n-1}, 1) \) is a vector of \( V \) but \(-1)(x_1, \ldots, x_{n-1}, 1) = (-x_1, \ldots, -x_{n-1}, -1)\) and this vector is not in \( V \) because the number in the \( n^{th} \) slot is negative, not nonnegative. Thus \( V \) is not closed with respect to scalar multiplication so it cannot be a subspace.

Example 5.1.26 Let \( X \) be the functions defined on \([0,1]\) and let \( V \) be those functions which are polynomials of degree 3 or less. Is \( V \) a subspace?

The answer is yes because if you add two polynomials of degree 3 or less, you get a polynomial of degree three or less. If you multiply such a polynomial by a scalar, you get another such polynomial.

5.1.4 Linear Independence

Probably the most important concept in linear algebra is that of linear independence.

Definition 5.1.27 A set of vectors \( \{v_1, \ldots, v_p\} \) is \textit{linearly independent} if the vector equation,

\[
c_1v_1 + \cdots + c_pv_p = 0
\]

has only the \textit{trivial solution},

\[
c_1 = c_2 = \cdots = c_p = 0.
\]

Otherwise the set of vectors is said to be \textit{dependent}.

Theorem 5.1.28 A set of vectors \( \{x_1, \ldots, x_p\} \) is linearly independent if and only if none of the vectors can be obtained as a linear combination of the others.

Proof: Suppose first that \( \{x_1, \cdots, x_p\} \) is linearly independent. If \( x_k = \sum_{j \neq k} c_j x_j \), then

\[
0 = 1x_k + \sum_{j \neq k} (-c_j) x_j,
\]

a nontrivial linear combination, contrary to assumption. This shows that if the set is linearly independent, then none of the vectors is a linear combination of the others.
Now suppose no vector is a linear combination of the others. It is desired to show that \( \{x_1, \ldots, x_p\} \) is linearly independent so suppose it is not. Then there exist scalars, \( c_i \), not all zero such that
\[
\sum_{i=1}^{p} c_i x_i = 0.
\]
Say \( c_k \neq 0 \). Then you can solve for \( x_k \) as
\[
x_k = \sum_{j \neq k} \left( -\frac{c_j}{c_k} \right) x_j
\]
contrary to assumption. This proves the lemma.

Restating this theorem in terms of dependent sets is useful.

**Corollary 5.1.29** A set of vectors \( \{x_1, \ldots, x_p\} \) is linearly dependent if and only if one of the vectors can be obtained as a linear combination of the others.

You can consider this vector which is a linear combination of the other vectors as dependent on the others. However, the precise meaning of dependent and independent pertains to a set of vectors, not an individual vector.

**Example 5.1.30** Let \( X \) denote the vector space of all functions defined on \([0,1]\). Let \( f_0(x) = 1, f_1(x) = x, \) and \( f_2(x) = x^2. \) Is the set of vectors, \( \{f_0, f_1, f_2\} \) independent or dependent?

Suppose \( af_0 + bf_1 + cf_2 = 0. \) Here the 0 refers to the 0 function, that function which sends every \( x \) to 0. Thus
\[
a + bx + cx^2 = 0
\]
for all \( x \in [0,1] \). Then since this holds for all such \( x \), you could assign the value 0 to \( x \) and conclude that \( a = 0. \) Thus
\[
bx + cx^2 = 0.
\]
Now since this holds for every \( x \in [0,1], \) you could take the derivative of both sides and obtain
\[
b + 2cx = 0.
\]
Now assign \( x \) the value 0 and conclude that \( b = 0. \) Hence \( 2cx = 0. \) Let \( x = 1 \) and conclude \( c = 0 \) also. Therefore, \( \{f_0, f_1, f_2\} \) is linearly independent.

**Example 5.1.31** Consider the three vectors
\[
\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.
\]
Are the vectors dependent or independent?

To find whether they are independent, you must determine whether there are non zero solutions, \( x, y, \) and \( z \) to
\[
x \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + y \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.
\]
This is equivalent to finding whether there are solutions to the system
\[
x + 2y = 0
\]
\[
y + z = 0
\]
\[
x + z = 0
\]
The augmented matrix for this system of equations is
\[
\begin{pmatrix}
1 & 2 & 0 & | & 0 \\
0 & 1 & 1 & | & 0 \\
1 & 0 & 1 & | & 0
\end{pmatrix}.
\]

Taking \(-1\) times the top row and adding to the bottom,
\[
\begin{pmatrix}
1 & 2 & 0 & | & 0 \\
0 & 1 & 1 & | & 0 \\
0 & -2 & 1 & | & 0
\end{pmatrix}.
\]

Now taking 2 times the second row and adding to the bottom yields
\[
\begin{pmatrix}
1 & 2 & 0 & | & 0 \\
0 & 1 & 1 & | & 0 \\
0 & 0 & -1 & | & 0
\end{pmatrix}
\]

and so you can see at this point that the only solution to the system (5.9) is the solution, \(x = y = z = 0\). Thus the vectors are linearly independent.

The following is called the **exchange theorem**. It is the fundamental idea upon which every significant idea in linear algebra is based. In particular it is essential for the next section. The proof is presented later on Page 105. Here is the statement.

**Theorem 5.1.32 (Exchange Theorem)** Let \(\{x_1, \cdots, x_r\}\) be a linearly independent set of vectors such that each \(x_i\) is in \(\text{span}(y_1, \cdots, y_s)\). Then \(r \leq s\).

In words and with slightly less precision it says that a spanning set has at least as many vectors as a linearly independent set.

### 5.1.5 Basis And Dimension

**Example 5.1.33** Consider the three vectors,
\[
\begin{pmatrix}
1 \\
2 \\
3
\end{pmatrix}, \quad \begin{pmatrix}
1 \\
0 \\
-1
\end{pmatrix}, \quad \begin{pmatrix}
2 \\
2 \\
2
\end{pmatrix}
\]

in \(\mathbb{R}^3\). Are they dependent or independent?

These three vectors are dependent because the third is the sum of the first two. The significance of this is that you can throw out the third of these vectors and obtain a smaller list of vectors which has the same span as follows.

\[
\begin{align*}
a \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + b \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} + c \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} \\
= a \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + b \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} + c \left( \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \right)
\end{align*}
\]

which is in the span of the first two vectors.

This is like the situation with the span of two vectors in which one was a multiple of the other. The span of the two was the same as the span of one of them. Geometrically, this yielded a line rather than a plane even though two vectors were listed.
When you have a dependent set of vectors, you can always throw out some, obtaining a smaller list which has the same span as the original list of vectors. The concept of a basis is related to this.

**Definition 5.1.34** Let \( V \) be a vector space. A finite set of vectors, \( \{ x_1, \ldots, x_r \} \) is a **basis** for \( V \) if \( \text{span}(x_1, \ldots, x_r) = V \) and \( \{ x_1, \ldots, x_r \} \) is linearly independent.

**Example 5.1.35** Here are three vectors. Determine whether they are a basis for \( \mathbb{R}^3 \).

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

First check for linear independence. If

\[
c_1 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + c_2 \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} + c_3 \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
\]

does it follow that \( c_1, c_2, c_3 \) are all equal to zero? In other words, is the only solution to the following system of equations the zero solution?

\[
\begin{align*}
c_1 + 2c_2 + 3c_3 &= 0 \\
c_1 + 0c_2 + c_3 &= 0 \\
c_1 + c_2 + 0c_3 &= 0
\end{align*}
\]

You know how to solve such systems by now. When you do so, you find the only solution is \( c_2 = 0, c_1 = 0, c_3 = 0 \). Therefore, these vectors are linearly independent.

Next check whether the vectors span \( \mathbb{R}^3 \). In other words, for any choice of \( x, y, z \), there must be constants, \( c_1, c_2, c_3 \) such that

\[
c_1 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + c_2 \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} + c_3 \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}
\]

This is equivalent to determining whether there exists a solution to the system of equations,

\[
\begin{align*}
c_1 + 2c_2 + 3c_3 &= x \\
c_1 + 0c_2 + c_3 &= y \\
c_1 + c_2 + 0c_3 &= z
\end{align*}
\]

for any choice of \( x, y, z \). In terms of matrices, this is equivalent to finding a solution to

\[
\begin{pmatrix}
1 & 2 & 3 \\
1 & 0 & 1 \\
1 & 1 & 0
\end{pmatrix}
\begin{pmatrix}
c_1 \\
c_2 \\
c_3
\end{pmatrix}
= \begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
\]

This can always be done in this case because the matrix on the left has an inverse. You know how to find the inverse of a matrix now. Its inverse is

\[
\begin{pmatrix}
-\frac{1}{4} & \frac{3}{4} & \frac{1}{2} \\
\frac{1}{4} & -\frac{3}{4} & \frac{1}{2} \\
\frac{1}{4} & \frac{1}{4} & -\frac{1}{2}
\end{pmatrix}
\]
Multiplying both sides by this inverse matrix you find the solution to above system of equations is

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} -\frac{1}{4} & \frac{3}{4} & \frac{1}{2} \\ \frac{1}{4} & -\frac{3}{4} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -\frac{1}{4}x + \frac{3}{4}y + \frac{1}{2}z \\ \frac{1}{4}x - \frac{3}{4}y + \frac{1}{2}z \\ \frac{1}{4}x + \frac{1}{4}y - \frac{1}{2}z \end{pmatrix}.$$

Since there exists such a solution, it follows the span of these vectors is the whole space and they are therefore a basis.

**Corollary 5.1.36** Let \{x_1, \ldots, x_r\} and \{y_1, \ldots, y_s\} be two bases\(^2\) of \(\mathbb{F}^n\). Then \(r = s = n\).

**Proof:** From the exchange theorem, \(r \leq s\) and \(s \leq r\). Now note the vectors,

$$e_i = (0, \ldots, 0, 1, 0 \cdots, 0) \quad \text{for } i = 1, 2, \ldots, n$$

are a basis for \(\mathbb{F}^n\). This proves the corollary.

There are many bases for \(\mathbb{F}^n\).

**Example 5.1.37** The vectors,

$$\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}$$

form a basis for \(\mathbb{R}^3\). So do the vectors

$$\begin{pmatrix} 1 \\ 7 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix}.$$  

You can verify this as in Example 5.1.35.

The following definition is actually included in the earlier definition of a basis but we list it here for the sake of emphasis.

**Definition 5.1.38** A finite set of vectors, \(\{x_1, \ldots, x_r\}\) is a basis for a subspace, \(V\) of \(\mathbb{F}^n\) if \(\text{span} (x_1, \ldots, x_r) = V\) and \(\{x_1, \ldots, x_r\}\) is linearly independent.

**Corollary 5.1.39** Let \(\{x_1, \ldots, x_r\}\) and \(\{y_1, \ldots, y_s\}\) be two bases for \(V\). Then \(r = s\).

**Proof:** From the exchange theorem, \(r \leq s\) and \(s \leq r\). Therefore, this proves the corollary.

**Definition 5.1.40** Let \(V\) be a subspace of \(\mathbb{F}^n\). Then \(\dim(V)\) read as the dimension of \(V\) is the number of vectors in a basis.

Of course you should wonder right now whether an arbitrary subspace of \(\mathbb{F}^n\) even has a basis. In fact it does and this is in the next theorem. First, here is an important lemma.

\(^2\)This is the plural form of basis. We could say basiss but it would involve an inordinate amount of hissing as in “The sixth sheik’s sixth sheep is sick”. This is the reason that bases is used instead of basiss.
5.1. VECTOR SPACES

Lemma 5.1.41 Suppose \( \mathbf{v} \notin \text{span} (\mathbf{u}_1, \ldots, \mathbf{u}_k) \) and \( \{\mathbf{u}_1, \ldots, \mathbf{u}_k, \mathbf{v}\} \) is linearly independent. Then \( \{\mathbf{u}_1, \ldots, \mathbf{u}_k, \mathbf{v}\} \) is also linearly independent.

Proof: Suppose \( \sum_{i=1}^{k} c_i \mathbf{u}_i + d \mathbf{v} = \mathbf{0} \). It is required to verify that each \( c_i = 0 \) and that \( d = 0 \). But if \( d \neq 0 \), then you can solve for \( \mathbf{v} \) as a linear combination of the vectors, \( \{\mathbf{u}_1, \ldots, \mathbf{u}_k\} \),

\[
\mathbf{v} = -\sum_{i=1}^{k} \left( \frac{c_i}{d} \right) \mathbf{u}_i
\]

c contrary to assumption. Therefore, \( d = 0 \). But then \( \sum_{i=1}^{k} c_i \mathbf{u}_i = \mathbf{0} \) and the linear independence of \( \{\mathbf{u}_1, \ldots, \mathbf{u}_k\} \) implies each \( c_i = 0 \) also. This proves the lemma.

Theorem 5.1.42 Let \( V \) be a nonzero subspace of \( \mathbb{F}^n \). Then \( V \) has a basis.

Proof: Let \( \mathbf{v}_1 \in V \) where \( \mathbf{v}_1 \neq \mathbf{0} \). If \( \text{span} \{\mathbf{v}_1\} = V \), stop. \( \{\mathbf{v}_1\} \) is a basis for \( V \). Otherwise, there exists \( \mathbf{v}_2 \in V \) which is not in \( \text{span} \{\mathbf{v}_1\} \). By Lemma 5.1.41 \( \{\mathbf{v}_1, \mathbf{v}_2\} \) is a linearly independent set of vectors. If \( \text{span} \{\mathbf{v}_1, \mathbf{v}_2\} = V \) stop, \( \{\mathbf{v}_1, \mathbf{v}_2\} \) is a basis for \( V \). If \( \text{span} \{\mathbf{v}_1, \mathbf{v}_2\} \neq V \), then there exists \( \mathbf{v}_3 \notin \text{span} \{\mathbf{v}_1, \mathbf{v}_2\} \) and \( \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\} \) is a larger linearly independent set of vectors. Continuing this way, the process must stop before \( n+1 \) steps because if not, it would be possible to obtain \( n+1 \) linearly independent vectors contrary to the exchange theorem. This proves the theorem.

Example 5.1.43 Consider the plane \( 2x + 3y + z = 0 \). Show this is a subspace and find a basis for the subspace.

When we write the equation, \( 2x + 3y + z = 0 \), we mean the set of all vectors, \( (x, y, z) \) such that \( 2x + 3y + z = 0 \). Why is this a subspace? Suppose \( \alpha, \beta \) are scalars and \( (x_1, y_1, z_1) \) and \( (x_2, y_2, z_2) \) are two vectors satisfying the condition determined by the equation. Does \( \alpha (x_1, y_1, z_1) + \beta (x_2, y_2, z_2) \) also satisfy the condition defined by the equation?

\[
\alpha (x_1, y_1, z_1) + \beta (x_2, y_2, z_2) = (\alpha x_1 + \beta x_2, \alpha y_1 + \beta y_2, \alpha z_1 + \beta z_2).
\]

\[
2 (\alpha x_1 + \beta x_2) + 3 (\alpha y_1 + \beta y_2) + (\alpha z_1 + \beta z_2)
\]

\[
= \alpha (2x_1 + 3y_1 + z_1) + \beta (2x_2 + 3y_2 + z_2) = \alpha 0 + \beta 0 = 0.
\]

Therefore, this does specify a subspace. It remains to find a basis for it.

From the equation, \( z = -2x - 3y \) and so the vectors which satisfy the equation are of the form

\[
\begin{pmatrix}
  x \\
  y \\
  -2x - 3y
\end{pmatrix}
= x
\begin{pmatrix}
  1 \\
  0 \\
  -2
\end{pmatrix}
+ y
\begin{pmatrix}
  0 \\
  1 \\
  -3
\end{pmatrix}, \quad (5.10)
\]

Therefore, a spanning set for this subspace is

\[
\left\{
\begin{pmatrix}
  1 \\
  0 \\
  -2
\end{pmatrix},
\begin{pmatrix}
  0 \\
  1 \\
  -3
\end{pmatrix}
\right\}.
\]

This will be a basis if it is linearly independent. Suppose

\[
c_1
\begin{pmatrix}
  1 \\
  0 \\
  -2
\end{pmatrix}
+ c_2
\begin{pmatrix}
  0 \\
  1 \\
  -3
\end{pmatrix}
= \begin{pmatrix}
  0 \\
  0 \\
  0
\end{pmatrix}.
\]

This is equivalent to the system of equations,
\[
\begin{align*}
c_1 + 0c_2 &= 0 \\
0c_1 + c_2 &= 0 \\
-2c_1 - 3c_2 &= 0
\end{align*}
\]

having augmented matrix,
\[
\begin{pmatrix}
1 & 0 & | & 0 \\
0 & 1 & | & 0 \\
-2 & -3 & | & 0
\end{pmatrix}
\]

and you see the only solution to this is \(c_1 = c_2 = 0\) from the top two lines of the augmented matrix. Therefore, these vectors form a basis for the subspace.

Are there any other bases? The answer is that there are infinitely many bases for this or any subspace. A simple way to see this is to replace one of the vectors by any nonzero multiple of itself. For example, the same subspace is obtained as the span of the two vectors,
\[
\left\{ \begin{pmatrix} 2 \\ 0 \\ -4 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -3 \end{pmatrix} \right\}.
\]

We just replaced the first vector by 2 times the first vector. A more interesting example is
\[
\left\{ \begin{pmatrix} 1 \\ 1 \\ -5 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -3 \end{pmatrix} \right\}.
\]

You can verify this is also a basis. We got it by replacing the first vector by the sum of the two. Another way to get a basis would be solve the equation for \(x\) rather than \(z\). Thus
\[
x = \frac{1}{2} (-z - 3y)
\]

and then vectors in the subspace are of the form
\[
\begin{pmatrix}
\frac{-1}{2}z - \frac{3}{2}y \\
y \\
z
\end{pmatrix}, y, z \in \mathbb{R}.
\]

Thus, in the same way as above, a basis is
\[
\left\{ \begin{pmatrix} \frac{-1}{2} \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} \frac{3}{2} \\ 1 \\ 0 \end{pmatrix} \right\}
\]

or if you like,
\[
\left\{ \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix} \right\}
\]

where we simply multiplied both vectors by 2.

In words the following corollary states that any linearly independent set of vectors can be enlarged to form a basis.

**Corollary 5.1.44** Let \(V\) be a subspace of \(\mathbb{R}^n\) 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 \(V\).
Proof: This follows immediately from the proof of Theorem 5.1.42. You do exactly the same argument except you start with \(\{v_1, \cdots, v_r\}\) rather than \(\{v_1\}\).

It is also true that any spanning set of vectors can be restricted to obtain a basis.

Theorem 5.1.45 Let \(V\) be a subspace of \(F^n\) 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 \(V\).

Proof: Let \(r\) be the smallest positive integer with the property that for some set, \(\{v_1, \cdots, v_r\} \subseteq \{u_1, \cdots, u_p\}\),

\[
\text{span} (v_1, \cdots, v_r) = V.
\]

Then \(r \leq p\) and it must be the case that \(\{v_1, \cdots, v_r\}\) is linearly independent because if it were not so, one of the vectors, say \(v_k\) would be a linear combination of the others. But then you could delete this vector from \(\{v_1, \cdots, v_r\}\) and the resulting list of \(r - 1\) vectors would still span \(V\) contrary to the definition of \(r\). This proves the theorem.

5.1.6 Proof Of Exchange Theorem

Theorem 5.1.46 (Exchange Theorem) Let \(\{x_1, \cdots, x_r\}\) be a linearly independent set of vectors such that each \(x_i\) is in \(\text{span}(y_1, \cdots, y_s)\). Then \(r \leq s\).

Proof: Define \(\text{span}\{y_1, \cdots, y_s\} \equiv V\), it follows there exist scalars, \(c_1, \cdots, c_s\) such that

\[
x_1 = \sum_{i=1}^{s} c_i y_i. \tag{5.11}
\]

Not all of these scalars can equal zero because if this were the case, it would follow that \(x_1 = 0\) and so \(\{x_1, \cdots, x_r\}\) would not be linearly independent. Indeed, if \(x_1 = 0\), \(1x_1 + \sum_{i=2}^{r} 0x_i = x_1 = 0\) and so there would exist a nontrivial linear combination of the vectors \(\{x_1, \cdots, x_r\}\) which equals zero.

Say \(c_k \neq 0\). Then solve (5.11)) for \(y_k\) and obtain

\[
y_k \in \text{span} \left( x_1, y_1, \cdots, y_{k-1}, y_{k+1}, \cdots, y_s \right).
\]

Define \(\{z_1, \cdots, z_{s-1}\}\) by

\[
\{z_1, \cdots, z_{s-1}\} \equiv \{y_1, \cdots, y_{k-1}, y_{k+1}, \cdots, y_s\}
\]

Therefore, \(\text{span}\{x_1, z_1, \cdots, z_{s-1}\} = V\) because if \(v \in V\), there exist constants \(c_1, \cdots, c_s\) such that

\[
v = \sum_{i=1}^{s-1} c_i z_i + c_s y_k.
\]

Now replace the \(y_k\) in the above with a linear combination of the vectors, \(\{x_1, z_1, \cdots, z_{s-1}\}\) to obtain \(v \in \text{span}\{x_1, z_1, \cdots, z_{s-1}\}\). The vector \(y_k\), in the list \(\{y_1, \cdots, y_s\}\), has now been replaced with the vector \(x_1\) and the resulting modified list of vectors has the same span as the original list of vectors, \(\{y_1, \cdots, y_s\}\).

Now suppose that \(r > s\) and that \(\text{span}\{x_1, \cdots, x_l, z_1, \cdots, z_p\} = V\) where the vectors, \(z_1, \cdots, z_p\) are each taken from the set, \(\{y_1, \cdots, y_s\}\) and \(l + p = s\). This has now been done for \(l = 1\) above. Then since \(r > s\), it follows that \(l \leq s < r\) and so \(l + 1 \leq r\). Therefore,
Let $x_{t+1}$ be a vector not in the list, $\{x_1, \ldots, x_t\}$ and since span $\{x_1, \ldots, x_t, z_1, \ldots, z_p\} = V$, there exist scalars, $c_i$ and $d_j$ such that

$$x_{t+1} = \sum_{i=1}^{t} c_i x_i + \sum_{j=1}^{p} d_j z_j.$$  \hfill (5.12)

Now not all the $d_j$ can equal zero because if this were so, it would follow that $\{x_1, \ldots, x_t\}$ would be a linearly dependent set because one of the vectors would equal a linear combination of the others. Therefore, (5.12) can be solved for one of the $z_i$, say $z_k$, in terms of $x_{t+1}$ and the other $z_i$ and just as in the above argument, replace that $z_i$ with $x_{t+1}$ to obtain

$$\text{span} \left\{ x_1, \ldots, x_t, x_{t+1}, z_1, \ldots, z_{k-1}, z_{k+1}, \ldots, z_p \right\} = V.$$  

Continue this way, eventually obtaining

$$\text{span} \{ x_1, \ldots, x_s \} = V.$$  

But then $x_r \in \text{span} \{ x_1, \ldots, x_s \}$ contrary to the assumption that $\{x_1, \ldots, x_r\}$ is linearly independent. Therefore, $r \leq s$ as claimed.

### 5.2 Exercises

1. Let $V$ denote the $2 \times 2$ matrices which are of the form $\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$ where $a$ and $b$ are numbers. Define addition and scalar multiplication in the usual way. Determine whether $V$ is a vector space.

2. Let $V$ denote the $2 \times 2$ matrices which are of the form $\begin{pmatrix} a & 1 \\ 0 & b \end{pmatrix}$ where $a$ and $b$ are numbers. Define addition and scalar multiplication in the usual way. Determine whether $V$ is a vector space.

3. Suppose you define addition of vectors in $\mathbb{R}^2$ in the following funny way. $(x, y) + (z, w) = (x + 2z, y + w)$ and scalar multiplication in the usual way, $k (x, y) = (kx, ky)$. With these operations, is $\mathbb{R}^2$ a vector space? Explain.

4. Suppose you define addition of vectors in $\mathbb{R}^2$ in the usual way. $(x, y) + (z, w) = (x + z, y + w)$ but scalar multiplication in the following strange way: $k (x, y) = (x, y)$. With these operations, is $\mathbb{R}^2$ a vector space? Explain.

5. Here are three vectors in $\mathbb{R}^3 : (1, 2, 1), (0, 1, 1), (1, 4, 3)$. Find a simple description of the span of these three vectors. Is the span of these three vectors a subspace?

6. Let $M = \{ u = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_3 = u_1 = 0 \}$. Is $M$ a subspace? Explain.

7. Let $M = \{ u = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_3 \geq u_1 \}$. Is $M$ a subspace? Explain.

8. Let $w \in \mathbb{R}^4$ and let $M = \{ u = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : \sum_{i=1}^{4} u_i w_i = 0 \}$. Is $M$ a subspace? Explain.

9. Let $M = \{ u = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_i \geq 0 \text{ for each } i = 1, 2, 3, 4 \}$. Is $M$ a subspace? Explain.
10. Let \( \mathbf{w}, \mathbf{w}_1 \) be given vectors in \( \mathbb{R}^4 \) and define
\[
M = \left\{ \mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : \sum_{i=1}^{4} u_i w_i = 0 \right. \left. \text{ and } \sum_{i=1}^{4} u_i w_{1i} = 0 \right\}.
\]
Is \( M \) a subspace? Explain.

11. Let \( M = \{ \mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : |u_1| \leq 4 \} \). Is \( M \) a subspace? Explain.

12. Let \( M = \{ \mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : \sin(u_1) = 1 \} \). Is \( M \) a subspace? Explain.

13. Here are three vectors. Determine whether they are linearly independent or linearly dependent.
\[
\begin{pmatrix} 1 \\ 2 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 2 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 3 \\ 0 \\ 0 \\ 0 \end{pmatrix}
\]

14. Verify that the set of real valued functions defined on \( D \) for some set, \( D \) is a vector space. See the examples listed after the definition of vector space.

15. Here are three vectors. Determine whether they are linearly independent or linearly dependent.
\[
\begin{pmatrix} 4 \\ 2 \\ 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 2 \\ 2 \\ 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 3 \\ 0 \\ 1 \\ 1 \end{pmatrix}
\]

16. Here are three vectors. Determine whether they are linearly independent or linearly dependent.
\[
\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 4 \\ 5 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix}
\]

17. Here are four vectors. Determine whether they span \( \mathbb{R}^3 \). Are these vectors linearly independent?
\[
\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 4 \\ 3 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}
\]

18. Here are four vectors. Determine whether they span \( \mathbb{R}^3 \). Are these vectors linearly independent?
\[
\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 4 \\ 3 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}
\]

19. Determine whether the following vectors are a basis for \( \mathbb{R}^3 \). If they are, explain why they are and if they are not, give a reason and tell whether they span \( \mathbb{R}^3 \).
\[
\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 4 \\ 3 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 2 \\ 4 \\ 0 \end{pmatrix}
\]
20. Determine whether the following vectors are a basis for \( \mathbb{R}^3 \). If they are, explain why they are and if they are not, give a reason and tell whether they span \( \mathbb{R}^3 \).

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

21. Determine whether the following vectors are a basis for \( \mathbb{R}^3 \). If they are, explain why they are and if they are not, give a reason and tell whether they span \( \mathbb{R}^3 \).

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

22. Determine whether the following vectors are a basis for \( \mathbb{R}^3 \). If they are, explain why they are and if they are not, give a reason and tell whether they span \( \mathbb{R}^3 \).

\[
\begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
\]

23. If you have 5 vectors in \( \mathbb{F}^5 \) and the vectors are linearly independent, can it always be concluded they span \( \mathbb{F}^5 \)? Explain.

24. If you have 6 vectors in \( \mathbb{F}^5 \), is it possible they are linearly independent? Explain.

25. Consider the vectors of the form

\[
\left\{ \begin{pmatrix} 2t + 3s \\ s - t \\ t + s \end{pmatrix} : s, t \in \mathbb{R} \right\}.
\]

Is this set of vectors a subspace of \( \mathbb{R}^3 \)? If so, explain why, give a basis for the subspace and find its dimension.

26. Consider the vectors of the form

\[
\left\{ \begin{pmatrix} 2t + 3s + u \\ s - t \\ t + s \\ u \end{pmatrix} : s, t, u \in \mathbb{R} \right\}.
\]

Is this set of vectors a subspace of \( \mathbb{R}^4 \)? If so, explain why, give a basis for the subspace and find its dimension.

27. Consider the vectors of the form

\[
\left\{ \begin{pmatrix} 2t + u \\ t + 3u \\ t + s + v \\ u \end{pmatrix} : s, t, u, v \in \mathbb{R} \right\}.
\]

Is this set of vectors a subspace of \( \mathbb{R}^4 \)? If so, explain why, give a basis for the subspace and find its dimension.
28. In any vector space, show that if \( x + y = 0 \), then \( y = -x \).

29. Show that in any vector space, \( 0x = 0 \). That is, the scalar 0 times the vector \( x \) gives the vector \( 0 \).

30. Show that in any vector space, \( (-1)x = -x \).

31. Let \( X \) be a vector space and suppose \( \{x_1, \ldots, x_k\} \) is a set of vectors from \( X \). Show that \( 0 \) is in \( \text{span}(x_1, \ldots, x_k) \).

32. Let the vectors be polynomials of degree no more than 3. Show that with the usual definitions of scalar multiplication wherein for \( p(x) \) a polynomial, \( (\alpha p)(x) = \alpha p(x) \) this is a vector space.

33. In the previous problem show that a basis for the vector space is \( \{1, x, x^2, x^3\} \).

34. Suppose \( A \) is an \( m \times n \) matrix. \( A(\mathbb{F}^n) \) is defined to be the set of vector which are equal to \( Ax \) for some \( x \in \mathbb{F}^n \). Show \( A(\mathbb{F}^n) \) is a subspace of \( \mathbb{F}^m \). If \( \{y_1, \ldots, y_p\} \) is a basis for \( A(\mathbb{F}^n) \), such that \( y_i = Ax_i \) where \( x_i \in \mathbb{F}^n \), show \( \{x_1, \ldots, x_p\} \) is linearly independent. Would the same conclusion hold if you only knew \( \{y_1, \ldots, y_p\} \) is a linearly independent set?

35. Suppose \( A \) is an \( m \times n \) matrix and \( \{x_1, \ldots, x_p\} \) is a linearly independent set in \( \mathbb{F}^n \), when can you conclude \( \{Ax_1, \ldots, Ax_p\} \) is a linearly independent set in \( \mathbb{F}^m \)?
Part II

Vector Calculus
Vectors And Points In $\mathbb{R}^n$

6.0.1 Outcomes

1. Evaluate the distance between two points in $\mathbb{R}^n$.

2. Be able to represent a vector in each of the following ways for $n = 2, 3$
   (a) as a directed arrow in $n$ space
   (b) as an ordered $n$ tuple
   (c) as a linear combination of unit coordinate vectors

3. Carry out the vector operations:
   (a) addition
   (b) scalar multiplication
   (c) find magnitude (norm or length)
   (d) Find the vector of unit length in the direction of a given vector.

4. Represent the operations of vector addition, scalar multiplication and norm geometrically.

5. Recall and apply the basic properties of vector addition, scalar multiplication and norm.


7. Describe an open ball in $\mathbb{R}^n$.

8. Determine whether a set in $\mathbb{R}^n$ is open, closed, or neither.

6.1 Distance in $\mathbb{R}^n$

How is distance between two points in $\mathbb{R}^n$ defined?

**Definition 6.1.1** Let $\mathbf{x} = (x_1, \cdots, x_n)$ and $\mathbf{y} = (y_1, \cdots, y_n)$ be two points in $\mathbb{R}^n$. Then $|\mathbf{x} - \mathbf{y}|$ to indicates the distance between these points and is defined as

$$
\text{distance between } \mathbf{x} \text{ and } \mathbf{y} \equiv |\mathbf{x} - \mathbf{y}| \equiv \left( \sum_{k=1}^{n} |x_k - y_k|^2 \right)^{1/2} .
$$
This is called the **distance formula**. The symbol, \( B(a, r) \) is defined by

\[
B(a, r) \equiv \{ x \in \mathbb{R}^n : |x - a| < r \}.
\]

This is called an **open ball** of radius \( r \) centered at \( a \). It gives all the points in \( \mathbb{R}^n \) which are closer to \( a \) than \( r \).

First of all note this is a generalization of the notion of distance in \( \mathbb{R} \). There the distance between two points, \( x \) and \( y \) was given by the absolute value of their difference. Thus \( |x - y| \) is equal to the distance between these two points on \( \mathbb{R} \). Now \( |x - y| = \left( (x - y)^2 \right)^{1/2} \) where the square root is always the positive square root. Thus it is the same formula as the above definition except there is only one term in the sum. Geometrically, this is the right way to define distance which is seen from the Pythagorean theorem. Consider the following picture in the case that \( n = 2 \).

There are two points in the plane whose Cartesian coordinates are \((x_1, x_2)\) and \((y_1, y_2)\) respectively. Then the solid line joining these two points is the hypotenuse of a right triangle which is half of the rectangle shown in dotted lines. What is its length? Note the lengths of the sides of this triangle are \(|y_1 - x_1|\) and \(|y_2 - x_2|\). Therefore, the Pythagorean theorem implies the length of the hypotenuse equals

\[
\left( |y_1 - x_1|^2 + |y_2 - x_2|^2 \right)^{1/2} = \left( (y_1 - x_1)^2 + (y_2 - x_2)^2 \right)^{1/2}
\]

which is just the formula for the distance given above.

Now suppose \( n = 3 \) and let \((x_1, x_2, x_3)\) and \((y_1, y_2, y_3)\) be two points in \( \mathbb{R}^3 \). Consider the following picture in which one of the solid lines joins the two points and a dotted line joins the points \((x_1, x_2, x_3)\) and \((y_1, y_2, x_3)\).
By the Pythagorean theorem, the length of the dotted line joining \((x_1, x_2, x_3)\) and \((y_1, y_2, x_3)\) equals
\[
\left( (y_1 - x_1)^2 + (y_2 - x_2)^2 \right)^{1/2}
\]
while the length of the line joining \((y_1, y_2, x_3)\) to \((y_1, y_2, y_3)\) is just \(|y_3 - x_3|\). Therefore, by the Pythagorean theorem again, the length of the line joining the points \((x_1, x_2, x_3)\) and \((y_1, y_2, y_3)\) equals
\[
\left\{ \left( (y_1 - x_1)^2 + (y_2 - x_2)^2 \right)^{1/2} + (y_3 - x_3)^2 \right\}^{1/2}
\]
which is again just the distance formula above.

This completes the argument that the above definition is reasonable. Of course you cannot continue drawing pictures in ever higher dimensions but there is not problem with the formula for distance in any number of dimensions. Here is an example.

**Example 6.1.2** Find the distance between the points in \(\mathbb{R}^4\), \(a = (1, 2, -4, 6)\) and \(b = (2, 3, -1, 0)\)

Use the distance formula and write
\[
|a - b|^2 = (1 - 2)^2 + (2 - 3)^2 + (-4 - (-1))^2 + (6 - 0)^2 = 47
\]
Therefore, \(|a - b| = \sqrt{47}\).

All this amounts to defining the distance between two points as the length of a straight line joining these two points. However, there is nothing sacred about using straight lines. One could define the distance to be the length of some other sort of line joining these points. It won’t be done in this book but sometimes this sort of thing is done.

Another convention which is usually followed, especially in \(\mathbb{R}^2\) and \(\mathbb{R}^3\) is to denote the first component of a point in \(\mathbb{R}^2\) by \(x\) and the second component by \(y\). In \(\mathbb{R}^3\) it is customary to denote the first and second components as just described while the third component is called \(z\).

**Example 6.1.3** Describe the points which are at the same distance between \((1, 2, 3)\) and \((0, 1, 2)\).

Let \((x, y, z)\) be such a point. Then
\[
\sqrt{(x - 1)^2 + (y - 2)^2 + (z - 3)^2} = \sqrt{x^2 + (y - 1)^2 + (z - 2)^2}.
\]

Squaring both sides
\[
(x - 1)^2 + (y - 2)^2 + (z - 3)^2 = x^2 + (y - 1)^2 + (z - 2)^2
\]
and so
\[
x^2 - 2x + 14 + y^2 - 4y + z^2 - 6z = x^2 + y^2 - 2y + 5 + z^2 - 4z
\]
which implies
\[
-2x + 14 - 4y - 6z = -2y + 5 - 4z
\]
and so
\[
2x + 2y + 2z = -9. \quad (6.1)
\]
Since these steps are reversible, the set of points which is at the same distance from the two given points consists of the points, \((x, y, z)\) such that (6.1) holds.

The following lemma is fundamental. It is a form of the Cauchy Schwarz inequality.
Lemma 6.1.4 Let \( x = (x_1, \cdots, x_n) \) and \( y = (y_1, \cdots, y_n) \) be two points in \( \mathbb{R}^n \). Then
\[
\left| \sum_{i=1}^{n} x_i y_i \right| \leq |x| |y|.
\] (6.2)

Proof: Let \( \theta \) be either 1 or \(-1\) such that
\[
\theta \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i (\theta y_i) = \left| \sum_{i=1}^{n} x_i y_i \right|
\]
and consider \( p(t) \equiv \sum_{i=1}^{n} (x_i + t\theta y_i)^2 \). Then for all \( t \in \mathbb{R} \),
\[
0 \leq p(t) = \sum_{i=1}^{n} x_i^2 + 2t \sum_{i=1}^{n} x_i \theta y_i + t^2 \sum_{i=1}^{n} y_i^2
\]
\[
= |x|^2 + 2t \sum_{i=1}^{n} x_i \theta y_i + t^2 |y|^2
\]
If \( |y| = 0 \) then (6.2) is obviously true because both sides equal zero. Therefore, assume \( |y| \neq 0 \) and then \( p(t) \) is a polynomial of degree two whose graph opens up. Therefore, it either has no zeroes, two zeroes or one repeated zero. If it has two zeroes, the above inequality must be violated because in this case the graph must dip below the \( x \) axis. Therefore, it either has no zeroes or exactly one. From the quadratic formula this happens exactly when
\[
4 \left( \sum_{i=1}^{n} x_i \theta y_i \right)^2 - 4 |x|^2 |y|^2 \leq 0
\]
and so
\[
\sum_{i=1}^{n} x_i \theta y_i = \left| \sum_{i=1}^{n} x_i y_i \right| \leq |x| |y|
\]
as claimed. This proves the inequality.

There are certain properties of the distance which are obvious. Two of them which follow directly from the definition are
\[
|x - y| = |y - x|
\]
\[
|x - y| \geq 0 \text{ and equals } 0 \text{ only if } y = x.
\]
The third fundamental property of distance is known as the triangle inequality. Recall that in any triangle the sum of the lengths of two sides is always at least as large as the third side. The following corollary is equivalent to this simple statement.

Corollary 6.1.5 Let \( x, y \) be points of \( \mathbb{R}^n \). Then
\[
|x + y| \leq |x| + |y|.
\]

Proof: Using the Cauchy Schwarz inequality, Lemma 6.1.4,
\[
|x + y|^2 \equiv \sum_{i=1}^{n} (x_i + y_i)^2
\]
\[
= \sum_{i=1}^{n} x_i^2 + 2 \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2
\]
\[
\leq |x|^2 + 2 |x| |y| \leq |y|^2
\]
\[
= (|x| + |y|)^2
\]
and so upon taking square roots of both sides,

\[ |x + y| \leq |x| + |y| \]

and this proves the corollary.

### 6.2 Open And Closed Sets

Eventually, one must consider functions which are defined on subsets of \( \mathbb{R}^n \) and their properties. The next definition will end up being quite important. It describe a type of subset of \( \mathbb{R}^n \) with the property that if \( x \) is in this set, then so is \( y \) whenever \( y \) is close enough to \( x \).

**Definition 6.2.1** Let \( U \subseteq \mathbb{R}^n \). \( U \) is an **open set** if whenever \( x \in U \), there exists \( r > 0 \) such that \( B(x, r) \subseteq U \). More generally, if \( U \) is any subset of \( \mathbb{R}^n \), \( x \in U \) is an **interior point** of \( U \) if there exists \( r > 0 \) such that \( x \in B(x, r) \subseteq U \). In other words \( U \) is an open set exactly when every point of \( U \) is an interior point of \( U \).

If there is something called an open set, surely there should be something called a closed set and here is the definition of one.

**Definition 6.2.2** A subset, \( C \), of \( \mathbb{R}^n \) is called a **closed set** if \( \mathbb{R}^n \setminus C \) is an open set. They symbol, \( \mathbb{R}^n \setminus C \) denotes everything in \( \mathbb{R}^n \) which is not in \( C \). It is also called the **complement** of \( C \). The symbol, \( S^C \) is a short way of writing \( \mathbb{R}^n \setminus S \).

To illustrate this definition, consider the following picture.

You see in this picture how the edges are dotted. This is because an open set, can not include the edges or the set would fail to be open. For example, consider what would happen if you picked a point out on the edge of \( U \) in the above picture. Every open ball centered at that point would have in it some points which are outside \( U \). Therefore, such a point would violate the above definition. You also see the edges of \( B(x, r) \) dotted suggesting that \( B(x, r) \) ought to be an open set. This is intuitively clear but does require a proof. This will be done in the next theorem and will give examples of open sets. Also, you can see that if \( x \) is close to the edge of \( U \), you might have to take \( r \) to be very small.

It is roughly the case that open sets don’t have their skins while closed sets do. Here is a picture of a closed set, \( C \).
Note that $x \notin C$ and since $\mathbb{R}^n \setminus C$ is open, there exists a ball, $B(x, r)$ contained entirely in $\mathbb{R}^n \setminus C$. If you look at $\mathbb{R}^n \setminus C$, what would be its skin? It can’t be in $\mathbb{R}^n \setminus C$ and so it must be in $C$. This is a rough heuristic explanation of what is going on with these definitions. Also note that $\mathbb{R}^n$ and $\emptyset$ are both open and closed. Here is why. If $x \in \emptyset$, then there must be a ball centered at $x$ which is also contained in $\emptyset$. This must be considered to be true because there is nothing in $\emptyset$ so there can be no example to show it false\(^1\). Therefore, from the definition, it follows $\emptyset$ is open. It is also closed because if $x \notin \emptyset$, then $B(x, 1)$ is also contained in $\mathbb{R}^n \setminus \emptyset = \mathbb{R}^n$. Therefore, $\emptyset$ is both open and closed. From this, it follows $\mathbb{R}^n$ is also both open and closed.

**Theorem 6.2.3** Let $x \in \mathbb{R}^n$ and let $r \geq 0$. Then $B(x, r)$ is an open set. Also,

$$D(x, r) \equiv \{y \in \mathbb{R}^n : |y - x| \leq r\}$$

is a closed set.

**Proof:** Suppose $y \in B(x, r)$. It is necessary to show there exists $r_1 > 0$ such that $B(y, r_1) \subseteq B(x, r)$. Define $r_1 \equiv r - |x - y|$. Then if $|z - y| < r_1$, it follows from the above triangle inequality that

$$|z - x| = |z - y + y - x| \leq |z - y| + |y - x| < r_1 + |y - x| = r - |x - y| + |y - x| = r.$$ \(\text{Note that if } r = 0 \text{ then } B(x, r) = \emptyset, \text{ the empty set. This is because if } y \in \mathbb{R}^n, |x - y| \geq 0 \text{ and so } y \notin B(x, 0). \text{ Since } \emptyset \text{ has no points in it, it must be open because every point in it, (There are none.) satisfies the desired property of being an interior point.}

Now suppose $y \notin D(x, r)$. Then $|x - y| > r$ and defining $\delta \equiv |x - y| - r$, it follows that if $z \in B(y, \delta)$, then by the triangle inequality,

$$|x - z| \geq |x - y| - |y - z| \geq |x - y| - \delta = |x - y| - (|x - y| - r) = r$$

and this shows that $B(y, \delta) \subseteq \mathbb{R}^n \setminus D(x, r)$. Since $y$ was an arbitrary point in $\mathbb{R}^n \setminus D(x, r)$, it follows $\mathbb{R}^n \setminus D(x, r)$ is an open set which shows from the definition that $D(x, r)$ is a closed set as claimed.

\(^1\)To a mathematician, the statement: Whenever a pig is born with wings it can fly must be taken as true. We do not consider biological or aerodynamic considerations in such statements. There is no such thing as a winged pig and therefore, all winged pigs must be superb flyers since there can be no example of one which is not. On the other hand we would also consider the statement: Whenever a pig is born with wings it can’t possibly fly, as equally true. The point is, you can say anything you want about the elements of the empty set and no one can gainsay your statement. Therefore, such statements are considered as true by default. You may say this is a very strange way of thinking about truth and ultimately this is because mathematics is not about truth. It is more about consistency and logic.
A picture which is descriptive of the conclusion of the above theorem which also implies the manner of proof is the following.

Recall \( \mathbb{R}^2 \) consists of ordered pairs, \( (x, y) \) such that \( x \in \mathbb{R} \) and \( y \in \mathbb{R} \). \( \mathbb{R}^2 \) is also written as \( \mathbb{R} \times \mathbb{R} \). In general, the following definition holds.

**Definition 6.2.4** The *Cartesian product* of two sets, \( A \times B \), means \( \{(a, b) : a \in A, b \in B\} \). If you have \( n \) sets, \( A_1, A_2, \ldots, A_n \)

\[
\prod_{i=1}^{n} A_i = \{ (x_1, x_2, \ldots, x_n) : \text{each } x_i \in A_i \}.
\]

Now suppose \( A \subseteq \mathbb{R}^m \) and \( B \subseteq \mathbb{R}^n \). Then if \( (x, y) \in A \times B \), \( x = (x_1, \ldots, x_m) \) and \( y = (y_1, \ldots, y_n) \), the following identification will be made.

\[
(x, y) = (x_1, \ldots, x_m, y_1, \ldots, y_n) \in \mathbb{R}^{m+n}.
\]

Similarly, starting with something in \( \mathbb{R}^{n+m} \), you can write it in the form \( (x, y) \) where \( x \in \mathbb{R}^m \) and \( y \in \mathbb{R}^n \). The following theorem has to do with the Cartesian product of two closed sets or two open sets. Also here is an important definition.

**Definition 6.2.5** A set, \( A \subseteq \mathbb{R}^n \) is said to be *bounded* if there exist finite intervals, \([a_i, b_i]\)

such that

\[
A \subseteq \prod_{i=1}^{n} [a_i, b_i].
\]

**Theorem 6.2.6** Let \( U \) be an open set in \( \mathbb{R}^m \) and let \( V \) be an open set in \( \mathbb{R}^n \). Then \( U \times V \) is an open set in \( \mathbb{R}^{m+n} \). If \( C \) is a closed set in \( \mathbb{R}^m \) and \( H \) is a closed set in \( \mathbb{R}^n \), then \( C \times H \) is a closed set in \( \mathbb{R}^{m+n} \). If \( C \) and \( H \) are bounded, then so is \( C \times H \).

**Proof:** Let \( (x, y) \in U \times V \). Since \( U \) is open, there exists \( r_1 > 0 \) such that \( B(x, r_1) \subseteq U \). Similarly, there exists \( r_2 > 0 \) such that \( B(y, r_2) \subseteq V \). Now

\[
B((x, y), \delta) \equiv \left\{ (s, t) \in \mathbb{R}^{m+n} : \sum_{k=1}^{m} |x_k - s_k|^2 + \sum_{j=1}^{n} |y_j - t_j|^2 < \delta^2 \right\}
\]

Therefore, if \( \delta \equiv \min(r_1, r_2) \) and \( (s, t) \in B((x, y), \delta) \), then it follows that \( s \in B(x, r_1) \subseteq U \) and that \( t \in B(y, r_2) \subseteq V \) which shows that \( B((x, y), \delta) \subseteq U \times V \). Hence \( U \times V \) is open as claimed.

Next suppose \( (x, y) \notin C \times H \). It is necessary to show there exists \( \delta > 0 \) such that \( B((x, y), \delta) \subseteq \mathbb{R}^{m+n} \setminus (C \times H) \). Either \( x \notin C \) or \( y \notin H \) since otherwise \( (x, y) \) would be a point of \( C \times H \). Suppose therefore, that \( x \notin C \). Since \( C \) is closed, there exists \( r > 0 \) such that \( B(x, r) \subseteq \mathbb{R}^m \setminus C \). Consider \( B((x, y), r) \). If \( (s, t) \in B((x, y), r) \), it follows that \( s \in B(x, r) \)
which is contained in $\mathbb{R}^m \setminus C$. Therefore, $B((x, y), r) \subseteq \mathbb{R}^{n+m} \setminus (C \times H)$ showing $C \times H$ is closed. A similar argument holds if $y \notin H$.

If $C$ is bounded, there exist $[a_i, b_i]$ such that $C \subseteq \prod_{i=1}^{m+n} [a_i, b_i]$ and if $H$ is bounded, $H \subseteq \prod_{i=m+1}^{m+n} [a_i, b_i]$ for intervals $[a_{m+1}, b_{m+1}], \ldots, [a_{m+n}, b_{m+n}]$. Therefore, $C \times H \subseteq \prod_{i=1}^{m+n} [a_i, b_i]$ and this establishes the last part of this theorem.

### 6.3 Exercises

1. Draw a picture of the points in $\mathbb{R}^2$ which are determined by the following ordered pairs.
   \begin{enumerate}
   \item (1, 2)
   \item (−2, −2)
   \item (−2, 3)
   \item (2, −5)
   \end{enumerate}

2. Does it make sense to write $(1, 2) + (2, 3, 1)$? Explain.

3. Draw a picture of the points in $\mathbb{R}^3$ which are determined by the following ordered triples.
   \begin{enumerate}
   \item (1, 2, 0)
   \item (−2, −2, 1)
   \item (−2, 3, −2)
   \end{enumerate}

4. You are given two points in $\mathbb{R}^3$, $(4, 5, −4)$ and $(2, 3, 0)$. Show the distance from the point, $(3, 4, −2)$ to the first of these points is the same as the distance from this point to the second of the original pair of points. Note that $3 = \frac{4+2}{2}, 4 = \frac{5+3}{2}$. Obtain a theorem which will be valid for general pairs of points, $(x, y, z)$ and $(x_1, y_1, z_1)$ and prove your theorem using the distance formula.

5. A sphere is the set of all points which are at a given distance from a single given point. Find an equation for the sphere which is the set of all points that are at a distance of 4 from the point $(1, 2, 3)$ in $\mathbb{R}^3$.

6. A parabola is the set of all points $(x, y)$ in the plane such that the distance from the point $(x, y)$ to a given point, $(x_0, y_0)$ equals the distance from $(x, y)$ to a given line. The point, $(x_0, y_0)$ is called the focus and the line is called the directrix. Find the equation of the parabola which results from the line $y = l$ and $(x_0, y_0)$ a given focus with $y_0 < l$. Repeat for $y_0 > l$.

7. A sphere centered at the point $(x_0, y_0, z_0)$ with radius $r$ consists of all points, $(x, y, z)$ whose distance to $(x_0, y_0, z_0)$ equals $r$. Write an equation for this sphere in $\mathbb{R}^3$.

8. Suppose the distance between $(x, y)$ and $(x', y')$ were defined to equal the larger of the two numbers $|x - x'|$ and $|y - y'|$. Draw a picture of the sphere centered at the point, $(0, 0, 0)$ if this notion of distance is used.

9. Repeat the same problem except this time let the distance between the two points be $|x - x'| + |y - y'|$. 


If \((x_1, y_1, z_1)\) and \((x_2, y_2, z_2)\) are two points such that \(|(x_i, y_i, z_i)| = 1\) for \(i = 1, 2\), show that in terms of the usual distance, \(|(x_1 + x_2, y_1 + y_2, z_1 + z_2)| < 1\). What would happen if you used the way of measuring distance given in Problem 8 \(|(x, y)| = \text{maximum of } |z|, |x|, |y|\)?

Give a simple description using the distance formula of the set of points which are at an equal distance between the two points \((x_1, y_1, z_1)\) and \((x_2, y_2, z_2)\).

Suppose you are given two points, \((-a, 0)\) and \((a, 0)\) in \(\mathbb{R}^2\) and a number, \(r > 2a\). The set of points described by

\[
\{(x, y) \in \mathbb{R}^2 : |(x, y) - (-a, 0)| + |(x, y) - (a, 0)| = r\}
\]

is known as an ellipse. The two given points are known as the focus points of the ellipse. Simplify this to the form \((\frac{x}{a})^2 + \left(\frac{y}{b}\right)^2 = 1\). This is a nice exercise in messy algebra.

Suppose you are given two points, \((-a, 0)\) and \((a, 0)\) in \(\mathbb{R}^2\) and a number, \(r > 2a\). The set of points described by

\[
\{(x, y) \in \mathbb{R}^2 : |(x, y) - (-a, 0)| - |(x, y) - (a, 0)| = r\}
\]

is known as hyperbola. The two given points are known as the focus points of the hyperbola. Simplify this to the form \((\frac{x}{a})^2 - \left(\frac{y}{b}\right)^2 = 1\). This is a nice exercise in messy algebra.

Let \((x_1, y_1)\) and \((x_2, y_2)\) be two points in \(\mathbb{R}^2\). Give a simple description using the distance formula of the perpendicular bisector of the line segment joining these two points. Thus you want all points, \((x, y)\) such that \(|(x, y) - (x_1, y_1)| = |(x, y) - (x_2, y_2)|\).

Let \(U = \{(x, y, z) \text{ such that } z > 0\}\). Determine whether \(U\) is open, closed or neither.

Let \(U = \{(x, y, z) \text{ such that } z \geq 0\}\). Determine whether \(U\) is open, closed or neither.

Let \(U = \{(x, y, z) \text{ such that } \sqrt{x^2 + y^2 + z^2} < 1\}\). Determine whether \(U\) is open, closed or neither.

Let \(U = \{(x, y, z) \text{ such that } \sqrt{x^2 + y^2 + z^2} \leq 1\}\). Determine whether \(U\) is open, closed or neither.

Show carefully that \(\mathbb{R}^n\) is both open and closed.

Show that every open set in \(\mathbb{R}^n\) is the union of open balls contained in it.

Show the intersection of any two open sets is an open set.

If \(S\) is a nonempty subset of \(\mathbb{R}^p\), a point, \(x\), is said to be a limit point of \(S\) if \(B(x, r)\) contains infinitely many points of \(S\) for each \(r > 0\). Show this is equivalent to saying that \(B(x, r)\) contains a point of \(S\) different than \(x\) for each \(r > 0\).

Closed sets were defined to be those sets which are complements of open sets. Show that a set is closed if and only if it contains all its limit points.
6.4 Physical Vectors

Suppose you push on something. What is important? There are really two things which are important, how hard you push and the direction you push. This illustrates the concept of force.

**Definition 6.4.1** *Force* is a vector. The magnitude of this vector is a measure of how hard it is pushing. It is measured in units such as Newtons or pounds or tons. Its direction is the direction in which the push is taking place.

Of course this is a little vague and will be left a little vague until the presentation of Newton’s second law later.

Vectors are used to model force and other physical vectors like velocity. What was just described would be called a force vector. It has two essential ingredients, its magnitude and its direction. Geometrically think of vectors as directed line segments or arrows as shown in the following picture in which all the directed line segments are considered to be the same vector because they have the same direction, the direction in which the arrows point, and the same magnitude (length).

Because of this fact that only direction and magnitude are important, it is always possible to put a vector in a certain particularly simple form. Let \( \overrightarrow{pq} \) be a directed line segment or vector. Then from Definition 1.3.4 it follows that \( \overrightarrow{pq} \) consists of the points of the form
\[
p + t(q - p)
\]
where \( t \in [0, 1] \). Subtract \( p \) from all these points to obtain the directed line segment consisting of the points
\[
0 + t(q - p), \ t \in [0, 1].
\]
The point in \( \mathbb{R}^n, q - p \), will represent the vector.

Geometrically, the arrow, \( \overrightarrow{pq} \), was slid so it points in the same direction and the base is at the origin, \( 0 \). For example, see the following picture.

In this way vectors can be identified with elements of \( \mathbb{R}^n \).

The magnitude of a vector determined by a directed line segment \( \overrightarrow{pq} \) is just the distance between the point \( p \) and the point \( q \). By the distance formula this equals
\[
\left( \sum_{k=1}^{n} (q_k - p_k)^2 \right)^{1/2} = |p - q|
\]
and for \( \mathbf{v} \) any vector in \( \mathbb{R}^n \) the magnitude of \( \mathbf{v} \) equals \( \left( \sum_{k=1}^{n} v_k^2 \right)^{1/2} = |\mathbf{v}|. \)

What is the geometric significance of scalar multiplication? If \( \mathbf{a} \) represents the vector, \( \mathbf{v} \) in the sense that when it is slid to place its tail at the origin, the element of \( \mathbb{R}^n \) at its point is \( \mathbf{a} \), what is \( r \mathbf{v} \)?

\[
|r \mathbf{v}| = \left( \sum_{k=1}^{n} (ra_k)^2 \right)^{1/2} = \left( \sum_{k=1}^{n} r^2 (a_k)^2 \right)^{1/2} = (r^2)^{1/2} \left( \sum_{k=1}^{n} a_k^2 \right)^{1/2} = |r| |\mathbf{v}|. 
\]

Thus the magnitude of \( r \mathbf{v} \) equals \( |r| \) times the magnitude of \( \mathbf{v} \). If \( r \) is positive, then the vector represented by \( r \mathbf{v} \) has the same direction as the vector, \( \mathbf{v} \) because multiplying by the scalar, \( r \), only has the effect of scaling all the distances. Thus the unit distance along any coordinate axis now has length \( r \) and in this rescaled system the vector is represented by \( \mathbf{a} \).

If \( r < 0 \) similar considerations apply except in this case all the \( a_i \) also change sign. From now on, \( \mathbf{a} \) will be referred to as a vector instead of an element of \( \mathbb{R}^n \) representing a vector as just described. The following picture illustrates the effect of scalar multiplication.

\[
\begin{align*}
\mathbf{v} & \rightarrow 2\mathbf{v} \\
\mathbf{v} & \rightarrow -2\mathbf{v}
\end{align*}
\]

Note there are \( n \) special vectors which point along the coordinate axes. These are

\[
\mathbf{e}_i \equiv (0, \ldots, 0, 1, 0, \ldots, 0)
\]

where the 1 is in the \( i^{th} \) slot and there are zeros in all the other spaces. See the picture in the case of \( \mathbb{R}^3 \).

The direction of \( \mathbf{e}_i \) is referred to as the \( i^{th} \) direction. Given a vector, \( \mathbf{v} = (a_1, \ldots, a_n) \), \( a_i \mathbf{e}_i \) is the \( i^{th} \) component of the vector. Thus \( a_i \mathbf{e}_i = (0, \cdots, 0, a_i, 0, \cdots, 0) \) and so this vector gives something possibly nonzero only in the \( i^{th} \) direction. Also, knowledge of the \( i^{th} \) component of the vector is equivalent to knowledge of the vector because it gives the entry in the \( i^{th} \) slot and for \( \mathbf{v} = (a_1, \cdots, a_n) \),

\[
\mathbf{v} = \sum_{k=1}^{n} a_k \mathbf{e}_k.
\]

What does addition of vectors mean physically? Suppose two forces are applied to some object. Each of these would be represented by a force vector and the two forces acting
VECTORS AND POINTS IN $\mathbb{R}^n$

together would yield an overall force acting on the object which would also be a force vector known as the resultant. Suppose the two vectors are $\mathbf{a} = \sum_{k=1}^{n} a_k \mathbf{e}_i$ and $\mathbf{b} = \sum_{k=1}^{n} b_k \mathbf{e}_i$. Then the vector, $\mathbf{a}$ involves a component in the $i^{th}$ direction, $a_i \mathbf{e}_i$ while the component in the $i^{th}$ direction of $\mathbf{b}$ is $b_i \mathbf{e}_i$. Then it seems physically reasonable that the resultant vector should have a component in the $i^{th}$ direction equal to $(a_i + b_i) \mathbf{e}_i$. This is exactly what is obtained when the vectors, $\mathbf{a}$ and $\mathbf{b}$ are added.

$$\mathbf{a} + \mathbf{b} = (a_1 + b_1, \cdots, a_n + b_n).$$

$$= \sum_{i=1}^{n} (a_i + b_i) \mathbf{e}_i.$$

Thus the addition of vectors according to the rules of addition in $\mathbb{R}^n$ which were presented earlier, yields the appropriate vector which duplicates the cumulative effect of all the vectors in the sum.

What is the geometric significance of vector addition? Suppose $\mathbf{u}$, $\mathbf{v}$ are vectors,

$$\mathbf{u} = (u_1, \cdots, u_n), \mathbf{v} = (v_1, \cdots, v_n)$$

Then $\mathbf{u} + \mathbf{v} = (u_1 + v_1, \cdots, u_n + v_n)$. How can one obtain this geometrically? Consider the directed line segment, $\overrightarrow{0u}$ and then, starting at the end of this directed line segment, follow the directed line segment $\overrightarrow{u(u + v)}$ to its end, $\mathbf{u} + \mathbf{v}$. In other words, place the vector $\mathbf{u}$ in standard position with its base at the origin and then slide the vector $\mathbf{v}$ till its base coincides with the point of $\mathbf{u}$. The point of this slid vector, determines $\mathbf{u} + \mathbf{v}$. To illustrate, see the following picture.

Note the vector $\mathbf{u} + \mathbf{v}$ is the diagonal of a parallelogram determined from the two vectors $\mathbf{u}$ and $\mathbf{v}$ and that identifying $\mathbf{u} + \mathbf{v}$ with the directed diagonal of the parallelogram determined by the vectors $\mathbf{u}$ and $\mathbf{v}$ amounts to the same thing as the above procedure.

An item of notation should be mentioned here. In the case of $\mathbb{R}^n$ where $n \leq 3$, it is standard notation to use $\mathbf{i}$ for $\mathbf{e}_1$, $\mathbf{j}$ for $\mathbf{e}_2$, and $\mathbf{k}$ for $\mathbf{e}_3$. Now here are some applications of vector addition to some problems.

**Example 6.4.2** There are three ropes attached to a car and three people pull on these ropes. The first exerts a force of $2\mathbf{i} + 3\mathbf{j} - 2\mathbf{k}$ Newtons, the second exerts a force of $3\mathbf{i} + 5\mathbf{j} + \mathbf{k}$ Newtons and the third exerts a force of $5\mathbf{i} - \mathbf{j} + 2\mathbf{k}$. Newtons. Find the total force in the direction of $\mathbf{i}$.

To find the total force add the vectors as described above. This gives $10\mathbf{i} + 7\mathbf{j} + \mathbf{k}$ Newtons. Therefore, the force in the $\mathbf{i}$ direction is 10 Newtons.

As mentioned earlier, the Newton is a unit of force like pounds.

**Example 6.4.3** An airplane flies North East at 100 miles per hour. Write this as a vector.

A picture of this situation follows.
The vector has length 100. Now using that vector as the hypotenuse of a right triangle having equal sides, the sides should be each of length \(100/\sqrt{2}\). Therefore, the vector would be \(100/\sqrt{2}i + 100/\sqrt{2}j\).

This example also motivates the concept of velocity.

Definition 6.4.4 The speed of an object is a measure of how fast it is going. It is measured in units of length per unit time. For example, miles per hour, kilometers per minute, feet per second. The velocity is a vector having the speed as the magnitude but also specifying the direction.

Thus the velocity vector in the above example is \(100/\sqrt{2}i + 100/\sqrt{2}j\).

Example 6.4.5 The velocity of an airplane is \(100i + j + k\) measured in kilometers per hour and at a certain instant of time its position is \((1, 2, 1)\). Here imagine a Cartesian coordinate system in which the third component is altitude and the first and second components are measured on a line from West to East and a line from South to North. Find the position of this airplane one minute later.

Consider the vector \((1, 2, 1)\), is the initial position vector of the airplane. As it moves, the position vector changes. After one minute the airplane has moved in the \(i\) direction a distance of \(100 \times \frac{1}{60} = \frac{5}{3}\) kilometer. In the \(j\) direction it has moved \(\frac{1}{60}\) kilometer during this same time, while it moves \(\frac{1}{60}\) kilometer in the \(k\) direction. Therefore, the new displacement vector for the airplane is

\[(1, 2, 1) + \left(\frac{5}{3}, \frac{1}{60}, \frac{1}{60}\right) = \left(\frac{8}{3}, \frac{121}{60}, \frac{121}{60}\right)\]

Example 6.4.6 A certain river is one half mile wide with a current flowing at 4 miles per hour from East to West. A man swims directly toward the opposite shore from the South bank of the river at a speed of 3 miles per hour. How far down the river does he find himself when he has swum across? How far does he end up swimming?

Consider the following picture.

You should write these vectors in terms of components. The velocity of the swimmer in still water would be \(3j\) while the velocity of the river would be \(-4i\). Therefore, the velocity of the swimmer is \(-4i + 3j\). Since the component of velocity in the direction across the river is 3, it follows the trip takes 1/6 hour or 10 minutes. The speed at which he travels is \(\sqrt{4^2 + 3^2} = 5\) miles per hour and so he travels \(5 \times \frac{1}{6} = \frac{5}{6}\) miles. Now to find the distance
downstream he finds himself, note that if \( x \) is this distance, \( x \) and \( 1/2 \) are two legs of a right triangle whose hypotenuse equals \( 5/6 \) miles. Therefore, by the Pythagorean theorem the distance downstream is
\[
\sqrt{(5/6)^2 - (1/2)^2} = \frac{2}{3} \text{ miles.}
\]

6.5 Exercises

1. The wind blows from West to East at a speed of 50 kilometers per hour and an airplane is heading North West at a speed of 300 Kilometers per hour. What is the velocity of the airplane relative to the ground? What is the component of this velocity in the direction North?

2. In the situation of Problem 1 how many degrees to the West of North should the airplane head in order to fly exactly North. What will be the speed of the airplane?

3. In the situation of 2 suppose the airplane uses 34 gallons of fuel every hour at that air speed and that it needs to fly North a distance of 600 miles. Will the airplane have enough fuel to arrive at its destination given that it has 63 gallons of fuel?

4. An airplane is flying due north at 150 miles per hour. A wind is pushing the airplane due east at 40 miles per hour. After 1 hour, the plane starts flying 30° East of North. Assuming the plane starts at \((0, 0)\), where is it after 2 hours? Let North be the direction of the positive \( y \) axis and let East be the direction of the positive \( x \) axis.

5. City A is located at the origin while city B is located at \((100, 200)\) where distances are in miles. An airplane flies at 300 miles per hour in still air. This airplane wants to fly from city A to city B but the wind is blowing in the direction of the positive \( y \) axis at a speed of 20 miles per hour. Find a unit vector such that if the plane heads in this direction, it will end up at city B having flown the shortest possible distance. How long will it take to get there?

6. A certain river is one half mile wide with a current flowing at 2 miles per hour from East to West. A man swims directly toward the opposite shore from the South bank of the river at a speed of 3 miles per hour. How far down the river does he find himself when he has swam across? How far does he end up swimming?

7. A certain river is one half mile wide with a current flowing at 2 miles per hour from East to West. A man can swim at 3 miles per hour in still water. In what direction should he swim in order to travel directly across the river? What would the answer to this problem be if the river flowed at 3 miles per hour and the man could swim only at the rate of 2 miles per hour?

8. Three forces are applied to a point which does not move. Two of the forces are \(2i + j + 3k\) Newtons and \(i - 3j + 2k\) Newtons. Find the third force.

9. The total force acting on an object is to be \(2i + j + k\) Newtons. A force of \(-i + j + k\) Newtons is being applied. What other force should be applied to achieve the desired total force?

10. A bird flies from its nest 5 km. in the direction 60° north of east where it stops to rest on a tree. It then flies 10 km. in the direction due southeast and lands atop a telephone pole. Place an \( xy \) coordinate system so that the origin is the bird’s nest, and the positive \( x \) axis points east and the positive \( y \) axis points north. Find the displacement vector from the nest to the telephone pole.
11. A car is stuck in the mud. There is a cable stretched tightly from this car to a tree which is 20 feet long. A person grasps the cable in the middle and pulls with a force of 100 pounds perpendicular to the stretched cable. The center of the cable moves two feet and remains still. What is the tension in the cable? The tension in the cable is the force exerted on this point by the part of the cable nearer the car as well as the force exerted on this point by the part of the cable nearer the tree.
VECTORS AND POINTS IN $\mathbb{R}^N$
Vector Products

7.0.1 Outcomes

1. Evaluate a dot product from the angle formula or the coordinate formula.
2. Interpret the dot product geometrically.
3. Evaluate the following using the dot product:
   (a) the angle between two vectors
   (b) the magnitude of a vector
   (c) the work done by a constant force on an object
4. Evaluate a cross product from the angle formula or the coordinate formula.
5. Interpret the cross product geometrically.
6. Evaluate the following using the cross product:
   (a) the area of a parallelogram
   (b) the area of a triangle
   (c) physical quantities such as the torque and angular velocity.
7. Find the volume of a parallelepiped using the box product.
8. Recall, apply and derive the algebraic properties of the dot and cross products.

7.1 The Dot Product

There are two ways of multiplying vectors which are of great importance in applications. The first of these is called the dot product, also called the scalar product and sometimes the inner product.

Definition 7.1.1 Let \( \mathbf{a}, \mathbf{b} \) be two vectors in \( \mathbb{R}^n \) define \( \mathbf{a} \cdot \mathbf{b} \) as

\[
\mathbf{a} \cdot \mathbf{b} \equiv \sum_{k=1}^{n} a_k b_k.
\]

With this definition, there are several important properties satisfied by the dot product. In the statement of these properties, \( \alpha \) and \( \beta \) will denote scalars and \( \mathbf{a}, \mathbf{b}, \mathbf{c} \) will denote vectors.
Proposition 7.1.2 The dot product satisfies the following properties.

\[ \mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a} \] (7.1)

\[ \mathbf{a} \cdot \mathbf{a} \geq 0 \text{ and equals zero if and only if } \mathbf{a} = \mathbf{0} \] (7.2)

\[(\alpha \mathbf{a} + \beta \mathbf{b}) \cdot \mathbf{c} = \alpha (\mathbf{a} \cdot \mathbf{c}) + \beta (\mathbf{b} \cdot \mathbf{c}) \] (7.3)

\[\mathbf{c} \cdot (\alpha \mathbf{a} + \beta \mathbf{b}) = \alpha (\mathbf{c} \cdot \mathbf{a}) + \beta (\mathbf{c} \cdot \mathbf{b}) \] (7.4)

\[|\mathbf{a}|^2 = \mathbf{a} \cdot \mathbf{a} \] (7.5)

You should verify these properties. Also be sure you understand that (7.4) follows from the first three and is therefore redundant. It is listed here for the sake of convenience.

Example 7.1.3 Find \((1, 2, 0, -1) \cdot (0, 1, 2, 3)\).

This equals \(0 + 2 + 0 + (-3) = -1\).

Example 7.1.4 Find the magnitude of \(\mathbf{a} = (2, 1, 4, 2)\). That is, find \(|\mathbf{a}|\).

This is \(\sqrt{(2, 1, 4, 2) \cdot (2, 1, 4, 2)} = 5\).

The dot product satisfies a fundamental inequality known as the Cauchy Schwartz inequality.

Theorem 7.1.5 The dot product satisfies the inequality

\[|\mathbf{a} \cdot \mathbf{b}| \leq |\mathbf{a}| |\mathbf{b}|.\] (7.6)

Furthermore equality is obtained if and only if one of \(\mathbf{a}\) or \(\mathbf{b}\) is a scalar multiple of the other.

Proof: First note that if \(\mathbf{b} = \mathbf{0}\) both sides of (7.6) equal zero and so the inequality holds in this case. Therefore, it will be assumed in what follows that \(\mathbf{b} \neq \mathbf{0}\).

Define a function of \(t \in \mathbb{R}\)

\[ f(t) = (\mathbf{a} + t\mathbf{b}) \cdot (\mathbf{a} + t\mathbf{b}). \]

Then by (7.2), \(f(t) \geq 0\) for all \(t \in \mathbb{R}\). Also from (7.3), (7.4), (7.1), and (7.5)

\[ f(t) = \mathbf{a} \cdot (\mathbf{a} + t\mathbf{b}) + t\mathbf{b} \cdot (\mathbf{a} + t\mathbf{b}) \]

\[= \mathbf{a} \cdot \mathbf{a} + t (\mathbf{a} \cdot \mathbf{b}) + t\mathbf{b} \cdot \mathbf{a} + t^2 \mathbf{b} \cdot \mathbf{b} \]

\[= |\mathbf{a}|^2 + 2t (\mathbf{a} \cdot \mathbf{b}) + |\mathbf{b}|^2 t^2. \]

Now

\[ f(t) = |\mathbf{b}|^2 \left( t^2 + 2 \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} + \frac{|\mathbf{a}|^2}{|\mathbf{b}|^2} \right) \]

\[= |\mathbf{b}|^2 \left( t^2 + 2 \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} + \left( \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} \right)^2 - \left( \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} \right)^2 + \frac{|\mathbf{a}|^2}{|\mathbf{b}|^2} \right) \]

\[= |\mathbf{b}|^2 \left( t^2 + \left( \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} \right)^2 + \left( \frac{|\mathbf{a}|^2}{|\mathbf{b}|^2} - \left( \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} \right)^2 \right) \right) \geq 0 \]
7.1. THE DOT PRODUCT

for all \( t \in \mathbb{R} \). In particular \( f(t) \geq 0 \) when \( t = -\left(\frac{a \cdot b}{|b|^2}\right) \) which implies

\[
\frac{|a|^2}{|b|^2} - \left(\frac{a \cdot b}{|b|^2}\right)^2 \geq 0. \tag{7.7}
\]

Multiplying both sides by \( |b|^4 \),

\[
|a|^2 |b|^2 \geq (a \cdot b)^2
\]

which yields (7.6).

From Theorem 7.1.5, equality holds in (7.6) whenever one of the vectors is a scalar multiple of the other. It only remains to verify this is the only way equality can occur. If either vector equals zero, then equality is obtained in (7.6) so it can be assumed both vectors are non zero and that equality is obtained in (7.7). This implies that \( f(t) = 0 \) when \( t = -\left(\frac{a \cdot b}{|b|^2}\right) \) and so from (7.2), it follows that for this value of \( t \), \( a + tb = 0 \) showing \( a = -tb \). This proves the theorem.

You should note that the entire argument was based only on the properties of the dot product listed in (7.1) - (7.5). This means that whenever something satisfies these properties, the Cauchy Schwartz inequality holds. There are many other instances of these properties besides vectors in \( \mathbb{R}^n \).

The Cauchy Schwartz inequality allows a proof of the triangle inequality for distances in \( \mathbb{R}^n \) in much the same way as the triangle inequality for the absolute value.

Theorem 7.1.6 (Triangle inequality) For \( a, b \in \mathbb{R}^n \)

\[
|a + b| \leq |a| + |b| \tag{7.8}
\]

and equality holds if and only if one of the vectors is a nonnegative scalar multiple of the other. Also

\[
||a| - |b|| \leq |a - b| \tag{7.9}
\]

Proof: By properties of the dot product and the Cauchy Schwartz inequality,

\[
|a + b|^2 = (a + b) \cdot (a + b)
= (a \cdot a) + (a \cdot b) + (b \cdot a) + (b \cdot b)
= |a|^2 + 2(a \cdot b) + |b|^2
\leq |a|^2 + 2|a| |b| + |b|^2
\leq |a|^2 + 2|a| |b| + |b|^2
= (|a| + |b|)^2.
\]

Taking square roots of both sides you obtain (7.8).

It remains to consider when equality occurs. If either vector equals zero, then that vector equals zero times the other vector and the claim about when equality occurs is verified. Therefore, it can be assumed both vectors are nonzero. To get equality in the second inequality above, Theorem 7.1.5 implies one of the vectors must be a multiple of the other. Say \( b = \alpha a \). If \( \alpha < 0 \) then equality cannot occur in the first inequality because in this case

\[
(a \cdot b) = \alpha |a|^2 < 0 < |\alpha| |a|^2 = |a \cdot b|
\]

Therefore, \( \alpha \geq 0 \).
To get the other form of the triangle inequality,

\[ a = a - b + b \]

so

\[ |a| = |a - b + b| \leq |a - b| + |b|. \]

Therefore,

\[ |a| - |b| \leq |a - b| \] \hspace{1cm} (7.10)

Similarly,

\[ |b| - |a| \leq |b - a| = |a - b|. \] \hspace{1cm} (7.11)

It follows from (7.10) and (7.11) that (7.9) holds. This is because \(|a| - |b|\) equals the left side of either (7.10) or (7.11) and either way, \(|a| - |b|\) \leq \(|a - b|\). This proves the theorem.

7.2 The Geometric Significance Of The Dot Product

7.2.1 The Angle Between Two Vectors

Given two vectors, \(a\) and \(b\), the included angle is the angle between these two vectors which is less than or equal to 180 degrees. The dot product can be used to determine the included angle between two vectors. To see how to do this, consider the following picture.

![Diagram of vectors](image)

By the law of cosines,

\[ |a - b|^2 = |a|^2 + |b|^2 - 2|a||b|\cos \theta. \]

Also from the properties of the dot product,

\[ |a - b|^2 = (a - b) \cdot (a - b) \]
\[ = |a|^2 + |b|^2 - 2a \cdot b \]

and so comparing the above two formulas,

\[ a \cdot b = |a||b|\cos \theta. \] \hspace{1cm} (7.12)

In words, the dot product of two vectors equals the product of the magnitude of the two vectors multiplied by the cosine of the included angle. Note this gives a geometric description of the dot product which does not depend explicitly on the coordinates of the vectors.

**Example 7.2.1** Find the angle between the vectors \(2i + j - k\) and \(3i + 4j + k\).
The dot product of these two vectors equals $6 + 4 - 1 = 9$ and the norms are $\sqrt{4 + 1 + 1} = \sqrt{6}$ and $\sqrt{9 + 16 + 1} = \sqrt{26}$. Therefore, from (7.12) the cosine of the included angle equals

$$
\cos \theta = \frac{9}{\sqrt{26} \sqrt{6}} = .72058
$$

Now the cosine is known, the angle can be determined by solving the equation, $\cos \theta = .72058$. This will involve using a calculator or a table of trigonometric functions. The answer is $\theta = .76616$ radians or in terms of degrees, $\theta = .76616 \times \frac{360}{\pi} = 43.898^\circ$. Recall how this last computation is done. Set up a proportion, $\frac{x}{.76616} = \frac{360}{\pi}$, because $360^\circ$ corresponds to $2\pi$ radians. However, in calculus, you should get used to thinking in terms of radians and not degrees. This is because all the important calculus formulas are defined in terms of radians.

**Example 7.2.2** Let $u, v$ be two vectors whose magnitudes are equal to 3 and 4 respectively and such that if they are placed in standard position with their tails at the origin, the angle between $u$ and the positive $x$ axis equals $30^\circ$ and the angle between $v$ and the positive $x$ axis is $-30^\circ$. Find $u \cdot v$.

From the geometric description of the dot product in (7.12)

$$
u \cdot v = 3 \times 4 \times \cos (60^\circ) = 3 \times 4 \times 1/2 = 6.
$$

**Observation 7.2.3** Two vectors are said to be perpendicular if the included angle is $\pi/2$ radians ($90^\circ$). You can tell if two nonzero vectors are perpendicular by simply taking their dot product. If the answer is zero, this means they are are perpendicular because $\cos \theta = 0$.

**Example 7.2.4** Determine whether the two vectors, $2i + j - k$ and $1i + 3j + 5k$ are perpendicular.

When you take this dot product you get $2 + 3 - 5 = 0$ and so these two are indeed perpendicular.

**Definition 7.2.5** When two lines intersect, the angle between the two lines is the smaller of the two angles determined.

**Example 7.2.6** Find the angle between the two lines, $(1, 2, 0) + t (1, 2, 3)$ and $(0, 4, -3) + t (-1, 2, -3)$.

These two lines intersect, when $t = 0$ in the first and $t = -1$ in the second. It is only a matter of finding the angle between the direction vectors. One angle determined is given by

$$
\cos \theta = \frac{-6}{14} = -\frac{3}{7}.
$$

(7.13)

We don’t want this angle because it is obtuse. The angle desired is the acute angle given by

$$
\cos \theta = \frac{3}{7}.
$$

It is obtained by using replacing one of the direction vectors with $-1$ times it.
7.2.2 Work And Projections

Our first application will be to the concept of work. The physical concept of work does not in any way correspond to the notion of work employed in ordinary conversation. For example, if you were to slide a 150 pound weight off a table which is three feet high and shuffle along the floor for 50 yards, sweating profusely and exerting all your strength to keep the weight from falling on your feet, keeping the height always three feet and then deposit this weight on another three foot high table, the physical concept of work would indicate that the force exerted by your arms did no work during this project even though the muscles in your hands and arms would likely be very tired. The reason for such an unusual definition is that even though your arms exerted considerable force on the weight, enough to keep it from falling, the direction of motion was at right angles to the force they exerted. The only part of a force which does work in the sense of physics is the component of the force in the direction of motion. The work is defined to be the magnitude of the component of this force times the distance over which it acts in the case where this component of force points in the direction of motion and \((-1)\) times the magnitude of this component times the distance in case the force tends to impede the motion. Thus the work done by a force on an object as the object moves from one point to another is a measure of the extent to which the force contributes to the motion. This is illustrated in the following picture in the case where the given force contributes to the motion.

In this picture the force, \(F\) is applied to an object which moves on the straight line from \(p_1\) to \(p_2\). There are two vectors shown, \(F_\parallel\) and \(F_\perp\) and the picture is intended to indicate that when you add these two vectors you get \(F\) while \(F_\parallel\) acts in the direction of motion and \(F_\perp\) acts perpendicular to the direction of motion. Only \(F_\parallel\) contributes to the work done by \(F\) on the object as it moves from \(p_1\) to \(p_2\). From trigonometry, you see the magnitude of \(F_\parallel\) should equal \(|F|\cos\theta\). Thus, since \(F_\parallel\) points in the direction of the vector from \(p_1\) to \(p_2\), the total work done should equal

\[
|F||p_1\rightarrow p_2|\cos\theta = |F||p_2 - p_1|\cos\theta
\]

If the included angle had been obtuse, then the work done by the force, \(F\) on the object would have been negative because in this case, the force tends to impede the motion from \(p_1\) to \(p_2\) but in this case, \(\cos\theta\) would also be negative and so it is still the case that the work done would be given by the above formula. Thus from the geometric description of the dot product given above, the work equals

\[
|F||p_2 - p_1|\cos\theta = F_\cdot (p_2 - p_1)
\]

This explains the following definition.

**Definition 7.2.7** Let \(F\) be a force acting on an object which moves from the point, \(p_1\) to the point \(p_2\). Then the work done on the object by the given force equals \(F_\cdot (p_2 - p_1)\).

The concept of writing a given vector, \(F\) in terms of two vectors, one which is parallel to a given vector, \(D\) and the other which is perpendicular can also be explained with no reliance on trigonometry, completely in terms of the algebraic properties of the dot product.
As before, this is mathematically more significant than any approach involving geometry or trigonometry because it extends to more interesting situations. This is done next.

**Theorem 7.2.8** Let \( \mathbf{F} \) and \( \mathbf{D} \) be nonzero vectors. Then there exist unique vectors \( \mathbf{F} || \) and \( \mathbf{F} \perp \) such that

\[
\mathbf{F} = \mathbf{F} || + \mathbf{F} \perp \tag{7.14}
\]

where \( \mathbf{F} || \) is a scalar multiple of \( \mathbf{D} \), also referred to as \( \text{proj}_\mathbf{D} (\mathbf{F}) \), and \( \mathbf{F} \perp \cdot \mathbf{D} = 0 \).

**Proof:** Suppose (7.14) and \( \mathbf{F} || = \alpha \mathbf{D} \). Taking the dot product of both sides with \( \mathbf{D} \) and using \( \mathbf{F} \perp \cdot \mathbf{D} = 0 \), this yields

\[
\mathbf{F} \cdot \mathbf{D} = \alpha |\mathbf{D}|^2
\]

which requires \( \alpha = \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2} \). Thus there can be no more than one vector, \( \mathbf{F} || \). It follows \( \mathbf{F} \perp \) must equal \( \mathbf{F} - \mathbf{F} || \). This verifies there can be no more than one choice for both \( \mathbf{F} || \) and \( \mathbf{F} \perp \).

Now let

\[
\mathbf{F} || \equiv \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2} \mathbf{D}
\]

and let

\[
\mathbf{F} \perp = \mathbf{F} - \mathbf{F} || = \mathbf{F} - \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2} \mathbf{D}
\]

Then \( \mathbf{F} || = \alpha \mathbf{D} \) where \( \alpha = \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2} \). It only remains to verify \( \mathbf{F} \perp \cdot \mathbf{D} = 0 \). But

\[
\mathbf{F} \perp \cdot \mathbf{D} = \mathbf{F} \cdot \mathbf{D} - \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2} \mathbf{D} \cdot \mathbf{D} = \mathbf{F} \cdot \mathbf{D} - \frac{\mathbf{F} \cdot \mathbf{D} |\mathbf{D}|^2}{|\mathbf{D}|^2} = \mathbf{F} \cdot \mathbf{D} - \mathbf{F} \cdot \mathbf{D} = 0.
\]

This proves the theorem.

**Example 7.2.9** Let \( \mathbf{F} = 2\mathbf{i} + 7\mathbf{j} - 3\mathbf{k} \) Newtons. Find the work done by this force in moving from the point \((1, 2, 3)\) to the point \((-9, -3, 4)\) along the straight line segment joining these points where distances are measured in meters.

According to the definition, this work is

\[
(2\mathbf{i} + 7\mathbf{j} - 3\mathbf{k}) \cdot (-10\mathbf{i} - 5\mathbf{j} + \mathbf{k}) = -20 + (-35) + (-3) = -58 \text{ Newton meters.}
\]

Note that if the force had been given in pounds and the distance had been given in feet, the units on the work would have been foot pounds. In general, work has units equal to units of a force times units of a length. Instead of writing Newton meter, people write joule because a joule is by definition a Newton meter. That word is pronounced “jewel” and it is the unit of work in the metric system of units. Also be sure you observe that the work done by the force can be negative as in the above example. In fact, work can be either positive, negative, or zero. You just have to do the computations to find out.

**Example 7.2.10** Find \( \text{proj}_\mathbf{u} (\mathbf{v}) \) if \( \mathbf{u} = 2\mathbf{i} + 3\mathbf{j} - 4\mathbf{k} \) and \( \mathbf{v} = \mathbf{i} - 2\mathbf{j} + \mathbf{k} \).
VECTOR PRODUCTS

From the above discussion in Theorem 7.2.8, this is just

\[
\frac{1}{4 + 9 + 16} (i - 2j + k) \cdot (2i + 3j - 4k) (2i + 3j - 4k)
\]

\[= \frac{-8}{29} (2i + 3j - 4k) = -\frac{16}{29} i - \frac{24}{29} j + \frac{32}{29} k.
\]

Example 7.2.11 Suppose \(a\), and \(b\) are vectors and \(b_\perp = b - \text{proj}_a (b)\). What is the magnitude of \(b_\perp\) in terms of the included angle?

\[
|b_\perp|^2 = (b - \text{proj}_a (b)) \cdot (b - \text{proj}_a (b))
\]

\[= \left( b - \frac{b \cdot a}{|a|^2} a \right) \cdot \left( b - \frac{b \cdot a}{|a|^2} a \right)
\]

\[= |b|^2 - 2 \frac{(b \cdot a)^2}{|a|^2} + \left( \frac{b \cdot a}{|a|^2} \right)^2 |a|^2
\]

\[= |b|^2 \left( 1 - \frac{(b \cdot a)^2}{|a|^2 |b|^2} \right)
\]

\[= |b|^2 \left( 1 - \cos^2 \theta \right) = |b|^2 \sin^2 \theta.
\]

where \(\theta\) is the included angle between \(a\) and \(b\) which is less than \(\pi\) radians. Therefore, taking square roots,

\[|b_\perp| = |b| \sin \theta.
\]

7.2.3 The Parabolic Mirror, An Application

When light is reflected the angle of incidence is always equal to the angle of reflection. This is illustrated in the following picture in which a ray of light reflects off something like a mirror.

![Diagram](image)

An interesting problem is to design a curved mirror which has the property that it will direct all rays of light coming from a long distance away (essentially parallel rays of light) to a single point. You might be interested in a reflecting telescope for example or some sort of scheme for achieving high temperatures by reflecting the rays of the sun to a small area. Turning things around, you could place a source of light at the single point and desire to have the mirror reflect this in a beam of light consisting of parallel rays. How can you design such a mirror?
It turns out this is pretty easy given the above techniques for finding the angle between vectors. Consider the following picture.

\[
\begin{align*}
\text{(0, } p) & \quad \text{Tangent to the curved mirror} \\
(x, y(x)) & \quad \text{Piece of the curved mirror}
\end{align*}
\]

It suffices to consider this in a plane for \( x > 0 \) and then let the mirror be obtained as a surface of revolution. In the above picture, let \((0, p)\) be the special point at which all the parallel rays of light will be directed. This is set up so the rays of light are parallel to the \( y \) axis. The two indicated angles will be equal and the equation of the indicated curve will be \( y = y(x) \) while the reflection is taking place at the point \((x, y(x))\) as shown. To say the two angles are equal is to say their cosines are equal. Thus from the above,

\[
\frac{(0, 1) \cdot (1, y'(x))}{\sqrt{1 + y'(x)^2}} = \frac{(-x, p - y) \cdot (-1, -y'(x))}{\sqrt{x^2 + (y - p)^2}}.
\]

This follows because the vectors forming the sides of one of the angles are \((0, 1)\) and \((1, y'(x))\) while the vectors forming the other angle are \((-x, p - y)\) and \((-1, -y'(x))\). Therefore, this yields the differential equation,

\[
y'(x) = \frac{-y'(x)(p - y) + x}{\sqrt{x^2 + (y - p)^2}}
\]

which is written more simply as

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

Now let \( y - p = xv \) so that \( y' = xv' + v \). Then in terms of \( v \) the differential equation is

\[
xv' = \frac{1}{\sqrt{1 + v^2} - v} - v.
\]

This reduces to

\[
\left(\frac{1}{\sqrt{1 + v^2} - v}\right) \frac{dv}{dx} = \frac{1}{x}.
\]

If \( G \in \int \left(\frac{1}{\sqrt{1 + v^2} - v}\right) \, dv \), then a solution to the differential equation is of the form

\[
G(v) - \ln x = C
\]

where \( C \) is a constant. This is because if you differentiate both sides with respect to \( x \),

\[
G'(v) \frac{dv}{dx} - \frac{1}{x} = \left(\frac{1}{\sqrt{1 + v^2} - v}\right) \frac{dv}{dx} \frac{1}{x} = 0.
\]
To find $G \in \int \left( \frac{1}{\sqrt{1+v^2}} - v \right) dv$, use a trig. substitution, $v = \tan \theta$. Then in terms of $\theta$, the antiderivative becomes

$$\int \left( \frac{1}{\sec \theta - \tan \theta} - \tan \theta \right) \sec^2 \theta d\theta = \int \sec \theta d\theta = \ln |\sec \theta + \tan \theta| + C.$$ 

Now in terms of $v$, this is

$$\ln \left( v + \sqrt{1 + v^2} \right) = \ln x + c.$$ 

There is no loss of generality in letting $c = \ln C$ because $\ln$ maps onto $\mathbb{R}$. Therefore, from laws of logarithms,

$$\ln \left| v + \sqrt{1 + v^2} \right| = \ln x + c = \ln x + \ln C = \ln Cx.$$ 

Therefore,

$$v + \sqrt{1 + v^2} = Cx$$

and so

$$\sqrt{1 + v^2} = Cx - v.$$ 

Now square both sides to get

$$1 + v^2 = C^2 x^2 + v^2 - 2Cx v$$

which shows

$$1 = C^2 x^2 - 2C \frac{y - p}{x} = C^2 x^2 - 2C (y - p).$$

Solving this for $y$ yields

$$y = \frac{C}{2} x^2 + \left( p - \frac{1}{2C} \right)$$

and for this to correspond to reflection as described above, it must be that $C > 0$. As described in an earlier section, this is just the equation of a parabola. Note it is possible to choose $C$ as desired adjusting the shape of the mirror.

7.2.4 The Dot Product And Distance In $\mathbb{C}^n$

It is necessary to give a generalization of the dot product for vectors in $\mathbb{C}^n$. This definition reduces to the usual one in the case the components of the vector are real.

**Definition 7.2.12** Let $x, y \in \mathbb{C}^n$. Thus $x = (x_1, \ldots, x_n)$ where each $x_k \in \mathbb{C}$ and a similar formula holding for $y$. Then the dot product of these two vectors is defined to be

$$x \cdot y \equiv \sum_j x_j \overline{y_j} \equiv x_1 \overline{y_1} + \cdots + x_n \overline{y_n}.$$ 

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. The reason for this is that when you take the dot 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

$$x \cdot x = \sum_j x_j \overline{x_j} = \sum_j |x_j|^2 \geq 0.$$
If you didn’t place a conjugate as in the above definition, things wouldn’t work out correctly. For example,

\[(1 + i)^2 + 2^2 = 4 + 2i\]

and this is not a positive number.

The following properties of the dot product follow immediately from the definition and you should verify each of them.

**Properties of the dot product:**

1. \( \mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u} \).
2. If \( a, b \) are numbers and \( \mathbf{u}, \mathbf{v}, \mathbf{z} \) are vectors then \((a \mathbf{u} + b \mathbf{v}) \cdot \mathbf{z} = a (\mathbf{u} \cdot \mathbf{z}) + b (\mathbf{v} \cdot \mathbf{z}).\)
3. \( \mathbf{u} \cdot \mathbf{u} \geq 0 \) and it equals 0 if and only if \( \mathbf{u} = \mathbf{0} \).

The norm is defined in the usual way.

**Definition 7.2.13** For \( \mathbf{x} \in \mathbb{C}^n \),

\[|\mathbf{x}| \equiv \left( \sum_{k=1}^{n} |x_k|^2 \right)^{1/2} = (\mathbf{x} \cdot \mathbf{x})^{1/2}\]

Here is a fundamental inequality called the **Cauchy Schwarz inequality** which is stated here in \( \mathbb{C}^n \). First here is a simple lemma.

**Lemma 7.2.14** If \( z \in \mathbb{C} \) there exists \( \theta \in \mathbb{C} \) such that \( \theta z = |z| \) and \( |\theta| = 1 \).

**Proof:** Let \( \theta = 1 \) if \( z = 0 \) and otherwise, let \( \theta = \frac{\overline{z}}{|z|} \). Recall that for \( z = x + iy, \overline{z} = x - iy \) and \( |\overline{z}| = |z|^2 \).

**Theorem 7.2.15** (Cauchy Schwarz) The following inequality holds for \( x_i \) and \( y_i \in \mathbb{C} \).

\[|(\mathbf{x} \cdot \mathbf{y})| = \left| \sum_{i=1}^{n} x_i \overline{y}_i \right| \leq \left( \sum_{i=1}^{n} |x_i|^2 \right)^{1/2} \left( \sum_{i=1}^{n} |y_i|^2 \right)^{1/2} = |\mathbf{x}| |\mathbf{y}| \quad (7.15)\]

**Proof:** Let \( \theta \in \mathbb{C} \) such that \( |\theta| = 1 \) and

\[\theta \sum_{i=1}^{n} x_i \overline{y}_i = \sum_{i=1}^{n} x_i \overline{y}_i \]

Thus

\[\theta \sum_{i=1}^{n} x_i \overline{y}_i = \sum_{i=1}^{n} x_i (\overline{\theta} y_i) = \left| \sum_{i=1}^{n} x_i \overline{y}_i \right|.\]

Consider \( p(t) \equiv \sum_{i=1}^{n} (x_i + t \overline{y}_i) \left( x_i + t \overline{y}_i \right) \) where \( t \in \mathbb{R} \).

\[
0 \leq p(t) = \sum_{i=1}^{n} |x_i|^2 + 2t \text{Re} \left( \theta \sum_{i=1}^{n} x_i \overline{y}_i \right) + t^2 \sum_{i=1}^{n} |y_i|^2
\]

\[
= |\mathbf{x}|^2 + 2t \left| \sum_{i=1}^{n} x_i \overline{y}_i \right| + t^2 |\mathbf{y}|^2
\]
If \(|y| = 0\) then (3.22) is obviously true because both sides equal zero. Therefore, assume \(|y| \neq 0\) and then \(p(t)\) is a polynomial of degree two whose graph opens up. Therefore, it either has no zeroes, two zeros or one repeated zero. If it has two zeros, the above inequality must be violated because in this case the graph must dip below the \(x\) axis. Therefore, it either has no zeros or exactly one. From the quadratic formula this happens exactly when
\[
4 \left| \sum_{i=1}^{n} x_i \overline{y}_i \right|^2 - 4 |x|^2 |y|^2 \leq 0
\]
and so
\[
\left| \sum_{i=1}^{n} x_i \overline{y}_i \right| \leq |x| |y|
\]
as claimed. This proves the inequality.

By analogy to the case of \(\mathbb{R}^n\), length or magnitude of vectors in \(\mathbb{C}^n\) can be defined.

**Definition 7.2.16** Let \(z \in \mathbb{C}^n\). Then \(|z| \equiv (z \cdot z)^{1/2}\).

**Theorem 7.2.17** For length defined in Definition 7.2.16, the following hold.

\[
|z| \geq 0 \text{ and } |z| = 0 \text{ if and only if } z = 0 \tag{7.16}
\]

If \(\alpha\) is a scalar, \(|\alpha z| = |\alpha| |z|\) \tag{7.17}

\[
|z + w| \leq |z| + |w|. \tag{7.18}
\]

**Proof:** The first two claims are left as exercises. To establish the third, you use the same argument which was used in \(\mathbb{R}^n\).

\[
|z + w|^2 = (z + w, z + w) = z \cdot z + w \cdot w + w \cdot z + z \cdot w \\
= |z|^2 + |w|^2 + 2 \text{Re} w \cdot z \\
\leq |z|^2 + |w|^2 + 2 |w \cdot z| \\
\leq |z|^2 + |w|^2 + 2 |w| |z| = (|z| + |w|)^2.
\]

All other considerations such as open and closed sets and the like are identical in this more general context with the corresponding definition in \(\mathbb{R}^n\). The main difference is that here the scalars are complex numbers.

**Definition 7.2.18** Suppose you have a vector space, \(V\) and for \(z, w \in V\) and \(\alpha\) a scalar

a norm is a way of measuring distance or magnitude which satisfies the properties (7.16) - (7.18). Thus a norm is something which does the following.

\[
||z|| \geq 0 \text{ and } ||z|| = 0 \text{ if and only if } z = 0 \tag{7.19}
\]

If \(\alpha\) is a scalar, \(||\alpha z|| = |\alpha||z||\) \tag{7.20}

\[
||z + w|| \leq ||z|| + ||w||. \tag{7.21}
\]

Here is is understood that for all \(z \in V, ||z|| \in [0, \infty)\).
7.3  Exercises

1. Use formula (7.12) to verify the Cauchy Schwartz inequality and to show that equality occurs if and only if one of the vectors is a scalar multiple of the other.

2. For $\mathbf{u}, \mathbf{v}$ vectors in $\mathbb{R}^3$, define the product, $\mathbf{u} \ast \mathbf{v} \equiv u_1v_1 + 2u_2v_3 + 3u_3v_3$. Show the axioms for a dot product all hold for this funny product. Prove $|\mathbf{u} \ast \mathbf{v}| \leq (|\mathbf{u}|^2 (\mathbf{v} \ast \mathbf{v})^{1/2})$.  
   **Hint:** Do not try to do this with methods from trigonometry.

3. Find the angle between the vectors $3\mathbf{i} - \mathbf{j} - \mathbf{k}$ and $\mathbf{i} + 4\mathbf{j} + 2\mathbf{k}$.

4. Find the angle between the vectors $\mathbf{i} - 2\mathbf{j} + \mathbf{k}$ and $\mathbf{i} + 2\mathbf{j} - 7\mathbf{k}$.

5. Find $\text{proj}_u (\mathbf{v})$ where $\mathbf{v} = (1, 0, -2)$ and $\mathbf{u} = (1, 2, 3)$.

6. Find $\text{proj}_u (\mathbf{v})$ where $\mathbf{v} = (1, 2, -2)$ and $\mathbf{u} = (1, 0, 3)$.

7. Find $\text{proj}_u (\mathbf{v})$ where $\mathbf{v} = (1, 2, -2, 1)$ and $\mathbf{u} = (1, 2, 3, 0)$.

8. Does it make sense to speak of $\text{proj}_0 (\mathbf{v})$?

9. If $\mathbf{F}$ is a force and $\mathbf{D}$ is a vector, show $\text{proj}_{\mathbf{D}} (\mathbf{F}) = (|\mathbf{F}| \cos \theta) \mathbf{u}$ where $\mathbf{u}$ is the unit vector in the direction of $\mathbf{D}$, $\mathbf{u} = \mathbf{D} / |\mathbf{D}|$ and $\theta$ is the included angle between the two vectors, $\mathbf{F}$ and $\mathbf{D}$. $|\mathbf{F}| \cos \theta$ is sometimes called the component of the force, $\mathbf{F}$ in the direction, $\mathbf{D}$.

10. A boy drags a sled for 100 feet along the ground by pulling on a rope which is 20 degrees from the horizontal with a force of 10 pounds. How much work does this force do?

11. A boy drags a sled for 200 feet along the ground by pulling on a rope which is 30 degrees from the horizontal with a force of 20 pounds. How much work does this force do?

12. How much work in Newton meters does it take to slide a crate 20 meters along a loading dock by pulling on it with a 200 Newton force at an angle of 30° from the horizontal?

13. An object moves 10 meters in the direction of $\mathbf{j}$. There are two forces acting on this object, $\mathbf{F}_1 = \mathbf{i} + \mathbf{j} + 2\mathbf{k}$, and $\mathbf{F}_2 = -5\mathbf{i} + 2\mathbf{j} - 6\mathbf{k}$. Find the total work done on the object by the two forces. **Hint:** You can take the work done by the resultant of the two forces or you can add the work done by each force.

14. An object moves 10 meters in the direction of $\mathbf{j} + \mathbf{i}$. There are two forces acting on this object, $\mathbf{F}_1 = \mathbf{i} + 2\mathbf{j} + 2\mathbf{k}$, and $\mathbf{F}_2 = 5\mathbf{i} + 2\mathbf{j} - 6\mathbf{k}$. Find the total work done on the object by the two forces. **Hint:** You can take the work done by the resultant of the two forces or you can add the work done by each force.

15. An object moves 20 meters in the direction of $\mathbf{k} + \mathbf{j}$. There are two forces acting on this object, $\mathbf{F}_1 = \mathbf{i} + \mathbf{j} + 2\mathbf{k}$, and $\mathbf{F}_2 = \mathbf{i} + 2\mathbf{j} - 6\mathbf{k}$. Find the total work done on the object by the two forces. **Hint:** You can take the work done by the resultant of the two forces or you can add the work done by each force.

16. If $\mathbf{a}, \mathbf{b},$ and $\mathbf{c}$ are vectors. Show that $(\mathbf{b} + \mathbf{c})_\perp = \mathbf{b}_\perp + \mathbf{c}_\perp$ where $\mathbf{b}_\perp = \mathbf{b} - \text{proj}_\mathbf{a} (\mathbf{b})$.

17. In the discussion of the reflecting mirror which directs all rays to a particular point, $(0, p)$. Show that for any choice of positive $C$ this point is the focus of the parabola and the directrix is $y = p - \frac{1}{C}$. 

18. Suppose you wanted to make a solar powered oven to cook food. Are there reasons for using a mirror which is not parabolic? Also describe how you would design a good flash light with a beam which does not spread out too quickly.

19. Find \((1, 2, 3, 4) \cdot (2, 0, 1, 3)\).

20. Show that \((\mathbf{a} \cdot \mathbf{b}) = \frac{1}{4} \left[ |\mathbf{a} + \mathbf{b}|^2 - |\mathbf{a} - \mathbf{b}|^2 \right].\)

21. Prove from the axioms of the dot product the parallelogram identity, \(|\mathbf{a} + \mathbf{b}|^2 + |\mathbf{a} - \mathbf{b}|^2 = 2|\mathbf{a}|^2 + 2|\mathbf{b}|^2\).

22. Let \(\mathbf{A}\) and \(\mathbf{B}\) be a real \(m \times n\) matrix and let \(\mathbf{x} \in \mathbb{R}^n\) and \(\mathbf{y} \in \mathbb{R}^m\). Show \((\mathbf{A}\mathbf{x}, \mathbf{y})_{\mathbb{R}^m} = (\mathbf{x}, \mathbf{A}^T\mathbf{y})_{\mathbb{R}^n}\) where \((\cdot, \cdot)_{\mathbb{R}^k}\) denotes the dot product in \(\mathbb{R}^k\). In the notation above, \(\mathbf{A}\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot \mathbf{A}^T\mathbf{y}\). Use the definition of matrix multiplication to do this.

23. Use the result of Problem 22 to verify directly that \((\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T\) without making any reference to subscripts.

24. Suppose \(f, g\) are two continuous functions defined on \([0, 1]\). Define \((f \cdot g) = \int_0^1 f(x) g(x) \, dx\). Show this dot product satisfies conditions (7.1) - (7.5). Explain why the Cauchy Schwarz inequality continues to hold in this context and state the Cauchy Schwarz inequality in terms of integrals.

7.4 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 7.4.1** 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 our 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 7.4.2** Let \( \mathbf{a} \) and \( \mathbf{b} \) be two vectors in \( \mathbb{R}^n \). 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 spanned by \( \mathbf{a} \) and \( \mathbf{b} \).

The cross product satisfies the following properties.

For \( \alpha \) a scalar,

\[
(\alpha \mathbf{a}) \times \mathbf{b} = \alpha (\mathbf{a} \times \mathbf{b}) = \mathbf{a} \times (\alpha \mathbf{b}),
\]

(7.23)

For \( \mathbf{a}, \mathbf{b}, \) and \( \mathbf{c} \) vectors, one obtains the distributive laws,

\[
\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c},
\]

(7.24)

\[
(\mathbf{b} + \mathbf{c}) \times \mathbf{a} = \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}.
\]

(7.25)

Formula (7.22) 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 (7.23) 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 \( \mathbf{a} (\mathbf{a} \times \mathbf{b}) \) and \( \mathbf{a} \times (\alpha \mathbf{b}) \). Using this yields equality in (7.23). 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 (7.22),

\[
(\mathbf{b} + \mathbf{c}) \times \mathbf{a} = -\mathbf{a} \times (\mathbf{b} + \mathbf{c})
\]

\[
= -(\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c})
\]

\[
= \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}.
\]

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 7.4.3** 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

\[
\mathbf{a} \times \mathbf{b} = (a_2 b_3 - a_3 b_2) \mathbf{i} + (a_3 b_1 - a_1 b_3) \mathbf{j} + (a_1 b_2 - a_2 b_1) \mathbf{k}.
\]

(7.26)
Proof: From the above table and the properties of the cross product listed,

\[(a_1 \mathbf{i} + a_2 \mathbf{j} + a_3 \mathbf{k}) \times (b_1 \mathbf{i} + b_2 \mathbf{j} + b_3 \mathbf{k}) =
\]

\[a_1 b_2 \mathbf{i} \times \mathbf{j} + a_1 b_3 \mathbf{i} \times \mathbf{k} + a_2 b_1 \mathbf{j} \times \mathbf{i} + a_2 b_3 \mathbf{j} \times \mathbf{k} + a_3 b_1 \mathbf{k} \times \mathbf{i} + a_3 b_2 \mathbf{k} \times \mathbf{j}
\]

\[= a_1 b_2 \mathbf{k} - a_1 b_3 \mathbf{j} - a_2 b_1 \mathbf{k} + a_2 b_3 \mathbf{i} + a_3 b_1 \mathbf{j} - a_3 b_2 \mathbf{i}
\]

\[= (a_2 b_3 - a_3 b_2) \mathbf{i} + (a_3 b_1 - a_1 b_3) \mathbf{j} + (a_1 b_2 - a_2 b_1) \mathbf{k}
\]

(7.27)

This proves the proposition.

It is probably impossible for most people to remember (7.26). Fortunately, there is a somewhat easier way to remember it.

\[\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}
\]

(7.28)

where you expand the determinant along the top row. This yields

\[= (a_2 b_3 - a_3 b_2) \mathbf{i} - (a_1 b_3 - a_3 b_1) \mathbf{j} + (a_1 b_2 - a_2 b_1) \mathbf{k}
\]

(7.29)

which is the same as (7.27).

Example 7.4.4 Find \((\mathbf{i} - \mathbf{j} + 2\mathbf{k}) \times (3\mathbf{i} - 2\mathbf{j} + \mathbf{k})\).

Use (7.28) to compute this.

\[\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & -1 & 2 \\ 3 & -2 & 1 \end{vmatrix} = \begin{vmatrix} -1 & 2 \\ -2 & 1 \\ 1 & -1 \end{vmatrix} \mathbf{i} + \begin{vmatrix} 1 & 2 \\ 3 & 1 \\ 1 & -2 \end{vmatrix} \mathbf{j} + \begin{vmatrix} 1 & -1 \\ 3 & -2 \end{vmatrix} \mathbf{k}
\]

\[= 3\mathbf{i} + 5\mathbf{j} + \mathbf{k}.
\]

Example 7.4.5 Find the area of the parallelogram determined by \((\mathbf{i} - \mathbf{j} + 2\mathbf{k})\) and \((3\mathbf{i} - 2\mathbf{j} + \mathbf{k})\). These are the same two vectors in Example 7.4.4.

From Example 7.4.4 and the geometric description of the cross product, the area is just the norm of the vector obtained in Example 7.4.4. Thus the area is \(\sqrt{9 + 25 + 1} = \sqrt{35}\).

Example 7.4.6 Find the area of the triangle determined by \((1, 2, 3), (0, 2, 5), \) and \((5, 1, 2)\).

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 \(\frac{1}{2} \sqrt{4 + 49 + 1} = \frac{3}{2} \sqrt{6}\).
7.4. THE CROSS PRODUCT

7.4.1 The Distributive Law For The Cross Product

This section gives a proof for (7.24), a fairly difficult topic. It is included here for the interested student. If you are satisfied with taking the distributive law on faith, it is not necessary to read this section. The proof given here is quite clever and follows the one given in [7]. Another approach, based on volumes of parallelepipeds is found in [24] and is discussed a little later.

Lemma 7.4.7 Let \( b \) and \( c \) be two vectors. Then \( b \times c = b \times c_\perp \) where \( c = c_\parallel + c_\perp \) and \( c_\perp \cdot b = 0 \).

Proof: Consider the following picture.

\[ c \perp \]
\[ \theta \]
\[ b \]

Now \( c_\perp = c - c_\parallel = c - \frac{b}{||b||}b \) and so \( c_\perp \) is in the plane determined by \( c \) and \( b \). Therefore, from the geometric definition of the cross product, \( b \times c \) and \( b \times c_\perp \) have the same direction. Now, referring to the picture,

\[
|b \times c_\perp| = |b||c_\perp| \\
= |b||c|\sin\theta \\
= |b \times c|.
\]

Therefore, \( b \times c \) and \( b \times c_\perp \) also have the same magnitude and so they are the same vector.

With this, the proof of the distributive law is in the following theorem.

Theorem 7.4.8 Let \( a, b, \) and \( c \) be vectors in \( \mathbb{R}^3 \). Then

\[
a \times (b + c) = a \times b + a \times c
\]  \hspace{1cm} (7.30)

Proof: Suppose first that \( a \cdot b = a \cdot c = 0 \). Now imagine \( a \) is a vector coming out of the page and let \( b, c \) and \( b + c \) be as shown in the following picture.

\[ a \times (b + c) \]
\[ a \times b \]
\[ a \times c \]
\[ b + c \]

Then \( a \times b, a \times (b + c), \) and \( a \times c \) are each vectors in the same plane, perpendicular to \( a \) as shown. Thus \( a \times c \cdot c = 0, a \times (b + c) \cdot (b + c) = 0, \) and \( a \times b \cdot b = 0 \). This implies that
to get \( \mathbf{a} \times \mathbf{b} \) you move counterclockwise through an angle of \( \pi/2 \) radians from the vector, \( \mathbf{b} \). Similar relationships exist between the vectors \( \mathbf{a} \times (\mathbf{b} + \mathbf{c}) \) and \( \mathbf{b} + \mathbf{c} \) and the vectors \( \mathbf{a} \times \mathbf{c} \) and \( \mathbf{c} \). Thus the angle between \( \mathbf{a} \times \mathbf{b} \) and \( \mathbf{a} \times (\mathbf{b} + \mathbf{c}) \) is the same as the angle between \( \mathbf{b} + \mathbf{c} \) and \( \mathbf{b} \) and the angle between \( \mathbf{a} \times \mathbf{c} \) and \( \mathbf{a} \times (\mathbf{b} + \mathbf{c}) \) is the same as the angle between \( \mathbf{c} \) and \( \mathbf{b} + \mathbf{c} \). In addition to this, since \( \mathbf{a} \) is perpendicular to these vectors, 

\[
|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}| \cdot |\mathbf{b}|, |\mathbf{a} \times (\mathbf{b} + \mathbf{c})| = |\mathbf{a}| \cdot |\mathbf{b} + \mathbf{c}|, \quad \text{and} \quad |\mathbf{a} \times \mathbf{c}| = |\mathbf{a}| \cdot |\mathbf{c}|.
\]

Therefore, 

\[
\frac{|\mathbf{a} \times (\mathbf{b} + \mathbf{c})|}{|\mathbf{b} + \mathbf{c}|} = \frac{|\mathbf{a} \times \mathbf{c}|}{|\mathbf{c}|} = \frac{|\mathbf{a} \times \mathbf{b}|}{|\mathbf{b}|} = |\mathbf{a}|
\]

and so 

\[
\frac{|\mathbf{a} \times (\mathbf{b} + \mathbf{c})|}{|\mathbf{a} \times \mathbf{c}|} = \frac{|\mathbf{b} + \mathbf{c}|}{|\mathbf{c}|}, \quad \frac{|\mathbf{a} \times (\mathbf{b} + \mathbf{c})|}{|\mathbf{a} \times \mathbf{b}|} = \frac{|\mathbf{b} + \mathbf{c}|}{|\mathbf{b}|}
\]

showing the triangles making up the parallelogram on the right and the four sided figure on the left in the above picture are similar. It follows the four sided figure on the left is in fact a parallelogram and this implies the diagonal is the vector sum of the vectors on the sides, yielding (7.30).

Now suppose it is not necessarily the case that \( \mathbf{a} \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{c} = 0 \). Then write \( \mathbf{b} = \mathbf{b}_\parallel + \mathbf{b}_\perp \) where \( \mathbf{b}_\perp \cdot \mathbf{a} = 0 \). Similarly \( \mathbf{c} = \mathbf{c}_\parallel + \mathbf{c}_\perp \). By the above lemma and what was just shown, 

\[
\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times (\mathbf{b} + \mathbf{c})_\perp = \mathbf{a} \times (\mathbf{b}_\perp + \mathbf{c}_\perp) = \mathbf{a} \times \mathbf{b}_\perp + \mathbf{a} \times \mathbf{c}_\perp = \mathbf{a} \times \mathbf{b}_\perp + \mathbf{a} \times \mathbf{c}_\perp = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}.
\]

This proves the theorem.

The result of Problem 16 of the exercises 7.3 is used to go from the first to the second line.

### 7.4.2 Torque

Imagine you are using a wrench to loosen a nut. The idea is to turn the nut by applying a force to the end of the wrench. If you push or pull the wrench directly toward or away from the nut, it should be obvious from experience that no progress will be made in turning the nut. The important thing is the component of force perpendicular to the wrench. It is this component of force which will cause the nut to turn. For example see the following picture.

In the picture a force, \( \mathbf{F} \) is applied at the end of a wrench represented by the position vector, \( \mathbf{R} \) and the angle between these two is \( \theta \). Then the tendency to turn will be \( |\mathbf{R}| \cdot |\mathbf{F}_\perp| = \)
\[ |\mathbf{R}| |\mathbf{F}| \sin \theta, \] which you recognize as the magnitude of the cross product of \( \mathbf{R} \) and \( \mathbf{F} \). If there were just one force acting at one point whose position vector is \( \mathbf{R} \), perhaps this would be sufficient, but what if there are numerous forces acting at many different points with neither the position vectors nor the force vectors in the same plane; what then? To keep track of this sort of thing, define for each \( \mathbf{R} \) and \( \mathbf{F} \), the torque vector,

\[ \tau \equiv \mathbf{R} \times \mathbf{F}. \]

This is also called the moment of the force, \( \mathbf{F} \). That way, if there are several forces acting at several points, the total torque can be obtained by simply adding up the torques associated with the different forces and positions.

**Example 7.4.9** Suppose \( \mathbf{R}_1 = 2\mathbf{i} - \mathbf{j} + 3\mathbf{k} \) meters and at the points determined by these vectors there are forces, \( \mathbf{F}_1 = \mathbf{i} - \mathbf{j} + 2\mathbf{k} \) and \( \mathbf{F}_2 = \mathbf{i} - 5\mathbf{j} + \mathbf{k} \) Newtons respectively. Find the total torque about the origin produced by these forces acting at the given points.

It is necessary to take \( \mathbf{R}_1 \times \mathbf{F}_1 + \mathbf{R}_2 \times \mathbf{F}_2 \). Thus the total torque equals

\[
\begin{vmatrix}
i & j & k \\
2 & -1 & 3 \\
1 & -1 & 2
\end{vmatrix} +
\begin{vmatrix}
i & j & k \\
1 & 2 & -6 \\
1 & -5 & 1
\end{vmatrix} = -27\mathbf{i} - 8\mathbf{j} - 8\mathbf{k} \text{ Newton meters}
\]

**Example 7.4.10** Find if possible a single force vector, \( \mathbf{F} \) which if applied at the point \( \mathbf{i} + \mathbf{j} + \mathbf{k} \) will produce the same torque as the above two forces acting at the given points.

This is fairly routine. The problem is to find \( \mathbf{F} = F_1\mathbf{i} + F_2\mathbf{j} + F_3\mathbf{k} \) which produces the above torque vector. Therefore,

\[
\begin{vmatrix}
i & j & k \\
1 & 1 & 1 \\
F_1 & F_2 & F_3
\end{vmatrix} = -27\mathbf{i} - 8\mathbf{j} - 8\mathbf{k}
\]

which reduces to \((F_3 - F_2)\mathbf{i} + (F_1 - F_3)\mathbf{j} + (F_2 - F_1)\mathbf{k} = -27\mathbf{i} - 8\mathbf{j} - 8\mathbf{k}\). This amounts to solving the system of three equations in three unknowns, \( F_1, F_2, \) and \( F_3, \)

\[
\begin{align*}
F_3 - F_2 &= -27 \\
F_1 - F_3 &= -8 \\
F_2 - F_1 &= -8
\end{align*}
\]

However, there is no solution to these three equations. (Why?) Therefore no single force acting at the point \( \mathbf{i} + \mathbf{j} + \mathbf{k} \) will produce the given torque.

### 7.4.3 Center Of Mass

The mass of an object is a measure of how much stuff there is in the object. An object has mass equal to one kilogram, a unit of mass in the metric system, if it would exactly balance a known one kilogram object when placed on a balance. The known object is one kilogram by definition. The mass of an object does not depend on where the balance is used. It would be one kilogram on the moon as well as on the earth. The weight of an object is something else. It is the force exerted on the object by gravity and has magnitude \( gm \) where \( g \) is a constant called the acceleration of gravity. Thus the weight of a one kilogram object would be different on the moon which has much less gravity, smaller \( g \), than on the earth. An important idea is that of the center of mass. This is the point at which an object will balance no matter how it is turned.
Definition 7.4.11 Let an object consist of \( p \) point masses, \( m_1, \ldots, m_p \) with the position of the \( k \)th of these at \( \mathbf{R}_k \). The center of mass of this object, \( \mathbf{R}_0 \) is the point satisfying
\[
\sum_{k=1}^{p} (\mathbf{R}_k - \mathbf{R}_0) \times g m_k \mathbf{u} = \mathbf{0}
\]
for all unit vectors, \( \mathbf{u} \).

The above definition indicates that no matter how the object is suspended, the total torque on it due to gravity is such that no rotation occurs. Using the properties of the cross product,
\[
\left( \sum_{k=1}^{p} \mathbf{R}_k g m_k - \mathbf{R}_0 \sum_{k=1}^{p} g m_k \right) \times \mathbf{u} = \mathbf{0}
\]
(7.31)
for any choice of unit vector, \( \mathbf{u} \). You should verify that if \( \mathbf{a} \times \mathbf{u} = \mathbf{0} \) for all \( \mathbf{u} \), then it must be the case that \( \mathbf{a} = \mathbf{0} \). Then the above formula requires that
\[
\sum_{k=1}^{p} \mathbf{R}_k g m_k - \mathbf{R}_0 \sum_{k=1}^{p} g m_k = \mathbf{0}.
\]
dividing by \( g \), and then by \( \sum_{k=1}^{p} m_k \),
\[
\mathbf{R}_0 = \frac{\sum_{k=1}^{p} \mathbf{R}_k m_k}{\sum_{k=1}^{p} m_k}.
\]
(7.32)
This is the formula for the center of mass of a collection of point masses. To consider the center of mass of a solid consisting of continuously distributed masses, you need the methods of calculus.

Example 7.4.12 Let \( m_1 = 5, m_2 = 6, \) and \( m_3 = 3 \) where the masses are in kilograms. Suppose \( m_1 \) is located at \( 2\mathbf{i} + 3\mathbf{j} + \mathbf{k} \), \( m_2 \) is located at \( \mathbf{i} - 3\mathbf{j} + 2\mathbf{k} \) and \( m_3 \) is located at \( 2\mathbf{i} - \mathbf{j} + 3\mathbf{k} \). Find the center of mass of these three masses.

Using (7.32)
\[
\mathbf{R}_0 = \frac{5(2\mathbf{i} + 3\mathbf{j} + \mathbf{k}) + 6(\mathbf{i} - 3\mathbf{j} + 2\mathbf{k}) + 3(2\mathbf{i} - \mathbf{j} + 3\mathbf{k})}{5 + 6 + 3} \\
= \frac{11}{7}\mathbf{i} - \frac{3}{7}\mathbf{j} + \frac{13}{7}\mathbf{k}
\]

7.4.4 Angular Velocity

Definition 7.4.13 In a rotating body, a vector, \( \Omega \) is called an angular velocity vector if the velocity of a point having position vector, \( \mathbf{u} \) relative to the body is given by \( \Omega \times \mathbf{u} \).

The existence of an angular velocity vector is the key to understanding motion in a moving system of coordinates. It is used to explain the motion on the surface of the rotating earth. For example, have you ever wondered why low pressure areas rotate counter clockwise in the upper hemisphere but clockwise in the lower hemisphere? To quantify these things, you will need the concept of an angular velocity vector. Details are presented later for interesting examples. Here we take a simple example. In the above example, think of a coordinate system fixed in the rotating body. Thus if you were riding on the rotating body, you would observe this coordinate system as fixed but it is not fixed.
7.4. THE CROSS PRODUCT

Example 7.4.14 A wheel rotates counter clockwise about the vector $i + j + k$ at 60 revolutions per minute. This means that if the thumb of your right hand were to point in the direction of $i + j + k$ your fingers of this hand would wrap in the direction of rotation. Find the angular velocity vector for this wheel. Assume the unit of distance is meters and the unit of time is minutes.

Let $\omega = 60 \times 2\pi = 120\pi$. This is the number of radians per minute corresponding to 60 revolutions per minute. Then the angular velocity vector is $\frac{120\pi}{\sqrt{3}}(i + j + k)$. This gives what you would expect in the case the position vector to the point is perpendicular to $i + j + k$ and at a distance of $r$. This is because of the geometric description of the cross product. The magnitude of the vector is $r \frac{120\pi}{\sqrt{3}}$ meters per minute and corresponds to the speed and an exercise with the right hand shows the direction is correct also. However, if this body is rigid, this will work for every other point in it, even those for which the position vector is not perpendicular to the given vector. A complete analysis of this is given later.

Example 7.4.15 A wheel rotates counter clockwise about the vector $i + j + k$ at 60 revolutions per minute exactly as in Example 7.4.14. Let $\{u_1, u_2, u_3\}$ denote an orthogonal right handed system attached to the rotating wheel in which $u_3 = \frac{1}{\sqrt{3}}(i + j + k)$. Thus $u_1$ and $u_2$ depend on time. Find the velocity of the point of the wheel located at the point $2u_1 + 3u_2 - u_3$.

Since $\{u_1, u_2, u_3\}$ is a right handed system like $i, j, k$, everything applies to this system in the same way as with $i, j, k$. Thus the cross product is given by

$$(au_1 + bu_2 + cu_3) \times (du_1 + eu_2 + fu_3)$$

Thus, in terms of the given vectors $u_i$, the angular velocity vector is

$$120\pi u_3$$

the velocity of the given point is

$$\begin{vmatrix} u_1 & u_2 & u_3 \\ 0 & 0 & 120\pi \\ 2 & 3 & -1 \end{vmatrix} = -360\pi u_1 + 240\pi u_2$$

in meters per minute. Note how this gives the answer in terms of these vectors which are fixed in the body, not in space. Since $u_i$ depends on $t$, this shows the answer in this case does also. Of course this is right. Just think of what is going on with the wheel rotating. Those vectors which are fixed in the wheel are moving in space. The velocity of a point in the wheel should be constantly changing. However, its speed will not change. The speed will be the magnitude of the velocity and this is

$$\sqrt{(-360\pi u_1 + 240\pi u_2) \cdot (-360\pi u_1 + 240\pi u_2)}$$

which from the properties of the dot product equals

$$\sqrt{(-360\pi)^2 + (240\pi)^2} = 120\sqrt{13}\pi$$

because the $u_i$ are given to be orthogonal.
7.4.5 \hspace{1em} \textbf{The Box Product}

\textbf{Definition 7.4.16} A parallelepiped determined by the three vectors, $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ consists of 
\[ \{ r \mathbf{a} + s \mathbf{b} + t \mathbf{c} : r, s, t \in [0,1] \}. \]

That is, if you pick three numbers, $r$, $s$, and $t$ each in $[0,1]$ and form $r \mathbf{a} + s \mathbf{b} + t \mathbf{c}$, 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.

\[ \begin{align*}
\mathbf{a} & \hspace{1em} \mathbf{c} \\
& \hspace{1em} \mathbf{a} \times \mathbf{c} \\
& \hspace{1em} \theta \\
\mathbf{b} & \hspace{1em} \end{align*} \]

You notice the area of the base of the parallelepiped, the parallelogram determined by the vectors, $\mathbf{a}$ and $\mathbf{c}$, has area equal to $|\mathbf{a} \times \mathbf{c}|$ while the altitude of the parallelepiped is $|\mathbf{b}| \cos \theta$ where $\theta$ is the angle shown in the picture between $\mathbf{b}$ and $\mathbf{a} \times \mathbf{c}$. Therefore, the volume of this parallelepiped is the area of the base times the altitude which is just
\[ |\mathbf{a} \times \mathbf{c}| |\mathbf{b}| \cos \theta = \mathbf{a} \times \mathbf{c} \cdot \mathbf{b}. \]

This expression is known as the box product and is sometimes written as $[\mathbf{a}, \mathbf{c}, \mathbf{b}]$. You should consider what happens if you interchange the $\mathbf{b}$ with the $\mathbf{c}$ or the $\mathbf{a}$ with the $\mathbf{c}$. 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.

\textbf{Example 7.4.17} Find the volume of the parallelepiped determined by the vectors, $\mathbf{i} + 2\mathbf{j} - 5\mathbf{k}, \mathbf{i} + 3\mathbf{j} - 6\mathbf{k}, 3\mathbf{i} + 2\mathbf{j} + 3\mathbf{k}$.

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.

\[ \begin{align*}
(i + 2j - 5k) \times (i + 3j - 6k) &= \begin{vmatrix} i & j & k \\ 1 & 2 & -5 \\ 1 & 3 & -6 \end{vmatrix} \\
&= 3i + j + k
\end{align*} \]

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 7.4.18 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.

An Alternate Proof Of The Distributive Law

Here is another proof of the distributive law for the cross product. Let \( \mathbf{x} \) be a vector. From the above observation,

\[
\mathbf{x} \cdot (\mathbf{a} \times (\mathbf{b} + \mathbf{c})) = (\mathbf{x} \times \mathbf{a}) \cdot (\mathbf{b} + \mathbf{c}) = (\mathbf{x} \times \mathbf{a}) \cdot \mathbf{b} + (\mathbf{x} \times \mathbf{a}) \cdot \mathbf{c} = \mathbf{x} \cdot (\mathbf{a} \times \mathbf{b}) + \mathbf{x} \cdot (\mathbf{a} \times \mathbf{c}).
\]

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 another way.

Observation 7.4.19 Suppose you have three vectors, \( \mathbf{u} = (a, b, c), \mathbf{v} = (d, e, f), \) and \( \mathbf{w} = (g, h, i) \). Then \( \mathbf{u} \cdot \mathbf{v} \times \mathbf{w} \) is given by the following.

\[
\mathbf{u} \cdot \mathbf{v} \times \mathbf{w} = \begin{vmatrix} i & j & k \\ d & e & f \\ g & h & i \end{vmatrix}
= \begin{vmatrix} e & f \\ h & i \end{vmatrix} - \begin{vmatrix} d & f \\ g & i \end{vmatrix} + \begin{vmatrix} d & e \\ g & h \end{vmatrix}
= \det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}.
\]

The message is that to take the box product, you can simply take the determinant of the matrix which results by letting the rows be the rectangular components of the given vectors in the order in which they occur in the box product.

7.5 Vector Identities And Notation

To begin with consider \( \mathbf{u} \times (\mathbf{v} \times \mathbf{w}) \) and it is desired to simplify this quantity. It turns out this is an important quantity which comes up in many different contexts. Let \( \mathbf{u} = (u_1, u_2, u_3) \) and let \( \mathbf{v} \) and \( \mathbf{w} \) be defined similarly.

\[
\mathbf{v} \times \mathbf{w} = \begin{vmatrix} i & j & k \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} = (v_2 w_3 - v_3 w_2) \mathbf{i} + (w_1 v_3 - v_1 w_3) \mathbf{j} + (v_1 w_2 - v_2 w_1) \mathbf{k}
\]

Next consider \( \mathbf{u} \times (\mathbf{v} \times \mathbf{w}) \) which is given by

\[
\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = \begin{vmatrix} i & j & k \\ u_1 & u_2 & u_3 \\ (v_2 w_3 - v_3 w_2) & (w_1 v_3 - v_1 w_3) & (v_1 w_2 - v_2 w_1) \end{vmatrix}.
\]
When you multiply this out, you get
\[
\begin{align*}
(i_1 u_2 v_2 + u_3 v_1 w_3 - w_1 u_2 v_2 - u_3 w_1 v_3) + j (v_2 u_1 w_1 + v_2 w_3 u_3 - w_2 u_1 v_1 - u_3 w_2 v_2) \\
+ k (u_1 w_1 v_3 + v_3 w_2 u_2 - u_1 v_3 w_3 - v_2 w_3 u_2)
\end{align*}
\]
and if you are clever, you see right away that
\[
(i_1 v_2 + j v_2 + k v_3) (u_1 w_1 + u_2 w_2 + u_3 w_3) - (i w_1 + j w_2 + k w_3) (u_1 v_1 + u_2 v_2 + u_3 v_3).
\]
Thus
\[
\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = \mathbf{v} (\mathbf{u} \cdot \mathbf{w}) - \mathbf{w} (\mathbf{u} \cdot \mathbf{v}).
\]
A related formula is
\[
(\mathbf{u} \times \mathbf{v}) \times \mathbf{w} = -[\mathbf{w} \times (\mathbf{u} \times \mathbf{v})] = -[\mathbf{u} (\mathbf{w} \cdot \mathbf{v}) - \mathbf{v} (\mathbf{w} \cdot \mathbf{u})] = \mathbf{v} (\mathbf{w} \cdot \mathbf{u}) - \mathbf{u} (\mathbf{w} \cdot \mathbf{v}).
\]
This derivation is simply wretched and it does nothing for other identities which may arise in applications. Actually, the above two formulas, (7.33) and (7.34) are sufficient for most applications if you are creative in using them, but there is another way. This other way allows you to discover such vector identities as the above without any creativity or any cleverness. Therefore, it is far superior to the above nasty computation. It is a vector identity discovering machine and it is this which is the main topic in what follows.

There are two special symbols, \(\delta_{ij}\) and \(\varepsilon_{ijk}\) which are very useful in dealing with vector identities. To begin with, here is the definition of these symbols.

**Definition 7.5.1** The symbol, \(\delta_{ij}\), called the Kronecker delta symbol is defined as follows.

\[
\delta_{ij} \equiv \begin{cases} 
1 & \text{if } i = j \\
0 & \text{if } i \neq j
\end{cases}
\]

With the Kronecker symbol, \(i\) and \(j\) can equal any integer in \(\{1, 2, \ldots, n\}\) for any \(n \in \mathbb{N}\).

**Definition 7.5.2** For \(i, j,\) and \(k\) integers in the set, \(\{1, 2, 3\}\), \(\varepsilon_{ijk}\) is defined as follows.

\[
\varepsilon_{ijk} \equiv \begin{cases} 
1 & \text{if } (i, j, k) \in (1, 2, 3), (2, 3, 1), \text{ or } (3, 1, 2) \\
-1 & \text{if } (i, j, k) \in (2, 1, 3), (1, 3, 2), \text{ or } (3, 2, 1) \\
0 & \text{if there are any repeated integers}
\end{cases}
\]

The subscripts \(ijk\) and \(ij\) in the above are called indices. A single one is called an index. This symbol, \(\varepsilon_{ijk}\) is also called the permutation symbol.

The way to think of \(\varepsilon_{ijk}\) is that \(\varepsilon_{123} = 1\) and if you switch any two of the numbers in the list \(i, j, k\), it changes the sign. Thus \(\varepsilon_{ijk} = -\varepsilon_{jik}\) and \(\varepsilon_{ijk} = -\varepsilon_{kji}\), etc. You should check that this rule reduces to the above definition. For example, it immediately implies that if there is a repeated index, the answer is zero. This follows because \(\varepsilon_{iij} = -\varepsilon_{iij}\) and so \(\varepsilon_{iij} = 0\).

It is useful to use the Einstein summation convention when dealing with these symbols. Simply stated, the convention is that you sum over the repeated index. Thus \(a_i b_i\) means \(\sum_i a_i b_i\). Also, \(\delta_{ij} x_j\) means \(\sum_j \delta_{ij} x_j = x_i\). When you use this convention, there is one very important thing to never forget. It is this: Never have an index be repeated more than once. Thus \(a_i b_i\) is all right but \(a_i b_i\) is not. The reason for this is that you end up getting confused about what is meant. If you want to write \(\sum_i a_i b_i c_i\), it is best to simply use the summation notation. There is a very important reduction identity connecting these two symbols.
Lemma 7.5.3 The following holds.

\[ \varepsilon_{ijk}\varepsilon_{irs} = (\delta_{jr}\delta_{ks} - \delta_{kr}\delta_{js}). \]

Proof: If \( \{j, k\} \neq \{r, s\} \) then every term in the sum on the left must have either \( \varepsilon_{ijk} \) or \( \varepsilon_{irs} \) contains a repeated index. Therefore, the left side equals zero. The right side also equals zero in this case. To see this, note that if the two sets are not equal, then there is one of the indices in one of the sets which is not in the other set. For example, it could be that \( j \) is not equal to either \( r \) or \( s \). Then the right side equals zero.

Therefore, it can be assumed \( \{j, k\} = \{r, s\} \). If \( i = r \) and \( j = s \) for \( s \neq r \), then there is exactly one term in the sum on the left and it equals 1. The right also reduces to 1 in this case. If there is a repeated index in \( \{j, k\} \), then every term in the sum on the left equals zero. The right also reduces to zero in this case because then \( j = k = r = s \) and so the right side becomes \((1)(1) - (-1)(-1) = 0\).

Proposition 7.5.4 Let \( u, v \) be vectors in \( \mathbb{R}^n \) where the Cartesian coordinates of \( u \) are \((u_1, \ldots, u_n)\) and the Cartesian coordinates of \( v \) are \((v_1, \ldots, v_n)\). Then \( u \cdot v = u_iv_i \). If \( u, v \) are vectors in \( \mathbb{R}^3 \), then

\[ (u \times v)_i = \varepsilon_{ijk}u_jv_k. \]

Also, \( \delta_{ik}a_k = a_i \).

Proof: The first claim is obvious from the definition of the dot product. The second is verified by simply checking it works. For example,

\[ u \times v \equiv \begin{vmatrix} i & j & k \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} \]

and so

\[ (u \times v)_1 = (u_2v_3 - u_3v_2). \]

From the above formula in the proposition,

\[ \varepsilon_{1jk}u_jv_k \equiv u_2v_3 - u_3v_2, \]

the same thing. The cases for \((u \times v)_2\) and \((u \times v)_3\) are verified similarly. The last claim follows directly from the definition.

With this notation, you can easily discover vector identities and simplify expressions which involve the cross product.

Example 7.5.5 Discover a formula which simplifies \((u \times v) \times w\).

From the above reduction formula,

\[ ((u \times v) \times w)_i = \varepsilon_{ijk}(u \times v)_j w_k = \varepsilon_{ijk}\varepsilon_{jrs}u_tv_sw_k = \varepsilon_{jik}\varepsilon_{jrs}u_tv_sw_k = -(\delta_{jr}\delta_{ks} - \delta_{kr}\delta_{js}) u_r v_s w_k = -u_iv_kw_k - u_kv_kw_k = u \cdot wv_i - v \cdot wu_i = ((u \cdot w)v - (v \cdot w)u)_i. \]
Since this holds for all \( i \), it follows that

\[
(u \times v) \times w = (u \cdot w) v - (v \cdot w) u.
\]

This is good notation and it will be used in the rest of the book whenever convenient.

### 7.6 Exercises

1. Show that if \( a \times u = 0 \) for all unit vectors, \( u \), then \( a = 0 \).

2. If you only assume (7.31) holds for \( u = i, j, k \), show that this implies (7.31) holds for all unit vectors, \( u \).

3. Let \( m_1 = 5, m_2 = 1, \) and \( m_3 = 4 \) where the masses are in kilograms and the distance is in meters. Suppose \( m_1 \) is located at \( 2i - 3j + k \), \( m_2 \) is located at \( i - 3j + 6k \) and \( m_3 \) is located at \( 2i + j + 3k \). Find the center of mass of these three masses.

4. Let \( m_1 = 2, m_2 = 3, \) and \( m_3 = 1 \) where the masses are in kilograms and the distance is in meters. Suppose \( m_1 \) is located at \( 2i - j + k \), \( m_2 \) is located at \( i - 2j + k \) and \( m_3 \) is located at \( 4i + j + 3k \). Find the center of mass of these three masses.

5. Find the angular velocity vector of a rigid body which rotates counter clockwise about the vector \( i - 2j + k \) at 40 revolutions per minute. Assume distance is measured in meters.

6. Let \( \{u_1, u_2, u_3\} \) be a right handed system with \( u_3 \) pointing in the direction of \( i - 2j + k \) and \( u_1 \) and \( u_2 \) being fixed with the body which is rotating at 40 revolutions per minute. Assuming all distances are in meters, find the constant speed of the point of the body located at \( 3u_1 + u_2 - u_3 \) in meters per minute.

7. Find the area of the triangle determined by the three points, \((1, 2, 3), (4, 2, 0)\) and \((-3, 2, 1)\).

8. Find the area of the triangle determined by the three points, \((1, 0, 3), (4, 1, 0)\) and \((-3, 1, 1)\).

9. Find the area of the triangle determined by the three points, \((1, 2, 3), (2, 3, 1)\) and \((0, 1, 2)\). Did something interesting happen here? What does it mean geometrically?

10. Find the area of the parallelogram determined by the vectors, \((1, 2, 3)\) and \((3, -2, 1)\).

11. Find the area of the parallelogram determined by the vectors, \((1, 0, 3)\) and \((4, -2, 1)\).

12. Find the area of the parallelogram determined by the vectors, \((1, -2, 2)\) and \((3, 1, 1)\).

13. Find the volume of the parallelepiped determined by the vectors, \(i - 7j - 5k, i - 2j - 6k, 3i + 2j + 3k\).

14. Find the volume of the parallelepiped determined by the vectors, \(i + j - 5k, i + 5j - 6k, 3i + j + 3k\).

15. Find the volume of the parallelepiped determined by the vectors, \(i + 6j + 5k, i + 5j - 6k, 3i + j + k\).

16. Suppose \(a, b,\) and \(c\) are three vectors whose components are all integers. Can you conclude the volume of the parallelepiped determined from these three vectors will always be an integer?
17. What does it mean geometrically if the box product of three vectors gives zero?

18. It is desired to find an equation of a plane containing the two vectors, \( \mathbf{a} \) and \( \mathbf{b} \) and the point \( \mathbf{0} \). Using Problem 17, show an equation for this plane is
\[
\begin{vmatrix}
  x & y & z \\
  a_1 & a_2 & a_3 \\
  b_1 & b_2 & b_3 \\
\end{vmatrix} = 0
\]
That is, the set of all \((x, y, z)\) such that the above expression equals zero.

19. Using the notion of the box product yielding either plus or minus the volume of the parallelepiped determined by the given three vectors, show that
\[
(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})
\]
In other words, the dot and the cross can be switched as long as the order of the vectors remains the same. **Hint:** There are two ways to do this, by the coordinate description of the dot and cross product and by geometric reasoning.

20. Is \( \mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) \times \mathbf{c} \)? What is the meaning of \( \mathbf{a} \times \mathbf{b} \times \mathbf{c} \)? Explain. **Hint:** Try \((\mathbf{i} \times \mathbf{j}) \times \mathbf{j}\).

21. Verify directly that the coordinate description of the cross product, \( \mathbf{a} \times \mathbf{b} \) has the property that it is perpendicular to both \( \mathbf{a} \) and \( \mathbf{b} \). Then show by direct computation that this coordinate description satisfies
\[
|\mathbf{a} \times \mathbf{b}|^2 = |\mathbf{a}|^2 |\mathbf{b}|^2 - (\mathbf{a} \cdot \mathbf{b})^2 = |\mathbf{a}|^2 |\mathbf{b}|^2 \left(1 - \cos^2(\theta)\right)
\]
where \(\theta\) is the angle included between the two vectors. Explain why \(|\mathbf{a} \times \mathbf{b}|\) has the correct magnitude. All that is missing is the material about the right hand rule. Verify directly from the coordinate description of the cross product that the right thing happens with regards to the vectors \(\mathbf{i}, \mathbf{j}, \mathbf{k}\). Next verify that the distributive law holds for the coordinate description of the cross product. This gives another way to approach the cross product. First define it in terms of coordinates and then get the geometric properties from this.

22. Discover a vector identity for \( \mathbf{u} \times (\mathbf{v} \times \mathbf{w}) \).

23. Discover a vector identity for \( (\mathbf{u} \times \mathbf{v}) \cdot (\mathbf{z} \times \mathbf{w}) \).

24. Discover a vector identity for \( (\mathbf{u} \times \mathbf{v}) \times (\mathbf{z} \times \mathbf{w}) \) in terms of box products.

25. Simplify \( (\mathbf{u} \times \mathbf{v}) \cdot (\mathbf{v} \times \mathbf{w}) \times (\mathbf{w} \times \mathbf{z}) \).

26. Simplify \( |\mathbf{u} \times \mathbf{v}|^2 + (\mathbf{u} \times \mathbf{v})^2 - |\mathbf{u}|^2 |\mathbf{v}|^2 \).

27. Prove that \( \varepsilon_{ijk} \varepsilon_{ijr} = 2 \delta_{kr} \).

28. If \( \mathbf{A} \) is a \( 3 \times 3 \) matrix such that \( \mathbf{A} = \begin{pmatrix} \mathbf{u} & \mathbf{v} & \mathbf{w} \end{pmatrix} \) where these are the columns of the matrix, \( \mathbf{A} \). Show that \(\det(\mathbf{A}) = \varepsilon_{ijk} u_i v_j w_k \).

29. If \( \mathbf{A} \) is a \( 3 \times 3 \) matrix, show \( \varepsilon_{rps} \det(\mathbf{A}) = \varepsilon_{ijk} A_{ri} A_{pj} A_{sk} \).
30. Suppose $A$ is a $3 \times 3$ matrix and $\det (A) \neq 0$. Show using 29 and 27 that

$$（A^{-1})_{ks} = \frac{1}{2 \det (A)} \varepsilon_{rps} \varepsilon_{ijk} A_{pj} A_{ri}.$$ 

31. When you have a rotating rigid body with angular velocity vector, $\Omega$ then the velocity, $u'$ is given by $u' = \Omega \times u$. It turns out that all the usual calculus rules such as the product rule hold. Also, $u''$ is the acceleration. Show using the product rule that for $\Omega$ a constant vector,

$$u'' = \Omega \times (\Omega \times u).$$

It turns out this is the centripetal acceleration. Note how it involves cross products. Things get really interesting when you move about on the rotating body. Weird forces are felt. This is in the section on moving coordinate systems.
Bases For $\mathbb{R}^n$

8.0.1 Outcomes
1. Recall and use the definition of an orthonormal basis.
2. Show that any orthonormal set of vectors is linearly independent.
3. Find an orthonormal basis using the Gram Schmidt process.
4. Define the dual basis for a given basis.
5. Find and define the metric tensor.
6. Find the dual basis for a given basis.
7. Explain how to use the metric tensor to write the dot product in terms of components with respect to a given basis which might not be orthonormal.

8.1 Orthonormal Bases
Not all bases for $\mathbb{F}^n$ are created equal. Recall $\mathbb{F}$ equals either $\mathbb{C}$ or $\mathbb{R}$ and the dot product is given by
\[ x \cdot y = \sum_j x_j y_j. \]
The best ones are orthonormal. Much of what follows will be for $\mathbb{F}^n$ in the interest of generality but you can substitute $\mathbb{R}$ for $\mathbb{F}$ if you like. Then later if you need it you can read it in full generality.

**Definition 8.1.1** Suppose $\{v_1, \ldots, v_k\}$ is a set of vectors in $\mathbb{F}^n$. It is an orthonormal set if $v_i \cdot v_j = \delta_{ij}$.

Every orthonormal set of vectors is automatically linearly independent.

**Proposition 8.1.2** Suppose $\{v_1, \ldots, v_k\}$ is an orthonormal set of vectors. Then it is linearly independent.

**Proof:** Suppose $c_i v_i = 0$ where summation takes place here and below over the repeated index. Then taking dot products with $v_j$, $0 = 0 \cdot v_j = c_i v_i \cdot v_j = c_i \delta_{ij} = c_j$. Since $j$ is arbitrary, this shows the set is linearly independent as claimed.

It turns out that if $X$ is any subspace of $\mathbb{F}^m$, then there exists an orthonormal basis for $X$. 

157
**Lemma 8.1.3** Let $X$ be a subspace of $\mathbb{F}^m$ 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 $(u_j, u_l) = \delta_{jl}$ and $\text{span}(x_1, \cdots, x_k) = \text{span}(u_1, \cdots, u_k)$. Then define

$$u_{k+1} = \frac{x_{k+1} - \sum_{j=1}^{k} (x_{k+1} \cdot u_j) u_j}{|x_{k+1} - \sum_{j=1}^{k} (x_{k+1} \cdot u_j) u_j|},$$

(8.1)

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 (8.1) 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 $l \leq k$,

$$(u_{k+1} \cdot u_l) = C \left( (x_{k+1} \cdot u_l) - \sum_{j=1}^{k} (x_{k+1} \cdot u_j) (u_j \cdot u_l) \right)$$

$$= C \left( (x_{k+1} \cdot u_l) - \sum_{j=1}^{k} (x_{k+1} \cdot u_j) \delta_{lj} \right)$$

$$= C ((x_{k+1} \cdot u_l) - (x_{k+1} \cdot u_l)) = 0.$$

The vectors, $\{u_j\}_{j=1}^n$, 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 Gram Schmidt process of the above lemma has major significance.

**Lemma 8.1.4** Let $A$ be an $m \times n$ matrix and let $A(\mathbb{F}^m)$ denote the set of vectors in $\mathbb{F}^m$ which are of the form $Ax$ for some $x \in \mathbb{F}^n$. Then $A(\mathbb{F}^m)$ is a subspace of $\mathbb{F}^m$.

**Proof:** Let $Ax$ and $Ay$ be two elements of $A(\mathbb{F}^m)$. It suffices to verify that if $a, b$ are scalars, then $aAx + bAy$ is also in $A(\mathbb{F}^m)$. But $aAx + bAy = A(ax + by)$ because $A$ is linear. This proves the lemma.

**Theorem 8.1.5** Let $y \in \mathbb{F}^m$ and let $A$ be an $m \times n$ matrix. Then there exists $x \in \mathbb{F}^n$ minimizing the function $|y - Ax|^2$. Furthermore, $x$ minimizes this function if and only if

$$(y - Ax) \cdot Aw = 0$$

for all $w \in \mathbb{F}^n$. 
8.1. ORTHONORMAL BASES

Proof: Let \( \{f_1, \ldots, f_r\} \) be an orthonormal basis for \( A(\mathbb{F}^n) \). Since \( A(\mathbb{F}^n) = \text{span} (f_1, \ldots, f_r) \), it follows that if you can find \( y_1, \ldots, y_r \) in such a way as to minimize

\[
\| y - \sum_{k=1}^{r} y_k f_k \|^2,
\]

then letting \( Ax = \sum_{k=1}^{r} y_k f_k \), it will follow that this \( x \) is the desired solution. Let \( y_1, \ldots, y_r \) be a list of scalars. Then from the definition of \( |\cdot| \) and the properties of the dot product,

\[
\| y - \sum_{k=1}^{r} y_k f_k \|^2 = \left( y - \sum_{k=1}^{r} y_k f_k \right) \cdot \left( y - \sum_{k=1}^{r} y_k f_k \right)
= |y|^2 - 2 \text{Re} \sum_{k=1}^{r} y_k \langle y, f_k \rangle + \sum_{k<l} y_k y_l \langle f_k, f_l \rangle
= |y|^2 - 2 \text{Re} \sum_{k=1}^{r} y_k \langle y, f_k \rangle + \sum_{k=1}^{r} |y_k|^2
= |y|^2 + \sum_{k=1}^{r} |y_k|^2 - 2 \text{Re} y_k \langle y, f_k \rangle
\]

Now complete the square to obtain

\[
= |y|^2 + \sum_{k=1}^{r} \left( |y_k|^2 - 2 \text{Re} y_k \langle y, f_k \rangle + |y \cdot f_k|^2 \right) - \sum_{k=1}^{r} |y \cdot f_k|^2
= |y|^2 + \sum_{k=1}^{r} |y_k - (y \cdot f_k)|^2 - \sum_{k=1}^{r} (y \cdot f_k)^2.
\]

This shows that the minimum is obtained when \( y_k = (y \cdot f_k) \). This proves the existence part of the Theorem.

To verify the other part, let \( t \in \mathbb{R} \) and consider

\[
|y - Ax + tw|^2 = \langle y - Ax - tAw, y - Ax - tAw \rangle
= |y - Ax|^2 - 2t \text{Re} \langle y - Ax, Aw \rangle + t^2 |Aw|^2.
\]

Then from the above equation, \( |y - Ax|^2 \leq |y - Az|^2 \) for all \( z \in \mathbb{F}^n \) if and only if for all \( w \in \mathbb{F}^n \) and \( t \in \mathbb{R} \)

\[
|y - Ax|^2 - 2t \text{Re} \langle y - Ax, Aw \rangle + t^2 |Aw|^2 \geq |y - Ax|^2
\]

and this happens if and only if for all \( t \in \mathbb{R} \) and \( w \in \mathbb{F}^n \),

\[
-2t \text{Re} \langle y - Ax, Aw \rangle + t^2 |Aw|^2 \geq 0,
\]

which occurs if and only if \( \text{Re} \langle y - Ax, Aw \rangle = 0 \) for all \( w \in \mathbb{F}^n \). (Why?)

This implies that \( \langle y - Ax, Aw \rangle = 0 \) for every \( w \in \mathbb{F}^n \) because there exists a complex number, \( \theta \) of magnitude 1 such that

\[
\langle y - Ax, Aw \rangle = \theta \langle y - Ax, Aw \rangle = \langle y - Ax, A\overline{\theta}w \rangle = \text{Re} \langle y - Ax, A\overline{\theta}w \rangle = 0.
\]


Definition 8.1.6 Let $A$ be an $m \times n$ matrix. Then
\[ A^* \equiv (A^T). \]
This means you take the transpose of $A$ and then replace each entry by its conjugate. This matrix is called the adjoint. Thus in the case of real matrices having only real entries, the adjoint is just the transpose.

Lemma 8.1.7 Let $A$ be an $m \times n$ matrix. Then
\[ Ax \cdot y = x \cdot A^*y \]

Proof: This follows from the definition. Using the repeated index summation convention,
\[
Ax \cdot y = A_{ij}x_jy_i \\
= x_jA^*_{ji}y_i \\
= x \cdot A^*y.
\]
This proves the lemma.

Corollary 8.1.8 A value of $x$ which solves the problem of Theorem 8.1.5 is obtained by solving the equation
\[ A^*Ax = A^*y \]
and furthermore, there exists a solution to this system of equations.

Proof: For $x$ the unique minimizer of Theorem 8.1.5, $(y - Ax) \cdot Aw = 0$ for all $w \in \mathbb{F}^n$ and from Lemma 8.1.7, this is the same as saying
\[ A^*(y - Ax) \cdot w = 0 \]
for all $w \in \mathbb{F}^n$. Therefore, there is a unique solution to the equation of this corollary and it solves the minimization problem of Theorem 8.1.5.

8.1.1 The Least Squares Regression Line

For the situation of the least squares regression line discussed here I will specialize to the case of $\mathbb{R}^n$ rather than $\mathbb{F}^n$ because it seems this case is by far the most interesting and the extra details are not justified by an increase in utility. Thus, everywhere you see $A^*$ it suffices to place $A^T$.

An important application of the above theorem is the problem of finding the least squares regression line in statistics. Suppose you are 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 $m, b$ to get as close as possible. The desired system is
\[
\begin{pmatrix}
y_1 \\
\vdots \\
y_n
\end{pmatrix} = \begin{pmatrix}
x_1 & 1 \\
\vdots & \vdots \\
x_n & 1
\end{pmatrix} \begin{pmatrix}
m \\
b
\end{pmatrix}
\]
which is of the form $y = Ax$ and it is desired to choose $m$ and $b$ to make
\[ \left| A \begin{pmatrix}
m \\
b
\end{pmatrix} - \begin{pmatrix}
y_1 \\
\vdots \\
y_n
\end{pmatrix} \right|^2 \]
as small as possible. According to Theorem 8.1.5 and Corollary 8.1.8, the best values for $m$ and $b$ occur as the solution to

$$AT\begin{pmatrix} m \\ b \end{pmatrix} = AT\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

where

$$A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}.$$ 

Thus, computing $AT\, A$,

$$\left( \sum_{i=1}^{n} x_i^2 \quad \sum_{i=1}^{n} x_i \quad n \right) \begin{pmatrix} m \\ b \end{pmatrix} = \left( \sum_{i=1}^{n} x_i y_i \right)$$

Solving this system of equations for $m$ and $b$,

$$m = -\frac{\left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right) + \left( \sum_{i=1}^{n} x_i y_i \right) n}{\left( \sum_{i=1}^{n} x_i^2 \right) n - \left( \sum_{i=1}^{n} x_i \right)^2}$$

and

$$b = -\frac{\left( \sum_{i=1}^{n} x_i \right) \sum_{i=1}^{n} x_i y_i + \left( \sum_{i=1}^{n} y_i \right) \sum_{i=1}^{n} x_i^2}{\left( \sum_{i=1}^{n} x_i^2 \right) n - \left( \sum_{i=1}^{n} x_i \right)^2}.$$ 

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 want to solve as well as possible for $a, b,$ and $c$ the system

$$\begin{pmatrix} x_1^2 & x_1 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

and one would use the same technique as above. Many other similar problems are important, including many in higher dimensions and they are all solved the same way.

8.1.2 The Fredholm Alternative

The next major result is called the Fredholm alternative. It comes from Theorem 8.1.5 and Lemma 8.1.7.

**Theorem 8.1.9** Let $A$ be an $m \times n$ matrix. Then there exists $x \in \mathbb{F}^n$ such that $Ax = y$ if and only if whenever $A^*z = 0$ it follows that $z \cdot y = 0$.

**Proof:** First suppose that for some $x \in \mathbb{F}^n$, $Ax = y$. Then letting $A^*z = 0$ and using Lemma 8.1.7

$$y \cdot z = Ax \cdot z = x \cdot A^*z = x \cdot 0 = 0.$$ 

This proves half the theorem.

To do the other half, suppose that whenever, $A^*z = 0$ it follows that $z \cdot y = 0$. It is necessary to show there exists $x \in \mathbb{F}^n$ such that $y = Ax$. From Theorem 8.1.5 there exists $x$ minimizing $\|y - Ax\|^2$ which therefore satisfies

$$(y - Ax) \cdot Aw = 0 \quad \text{(8.2)}$$
for all \( w \in F^n \). Therefore, for all \( w \in F^n \),
\[
A^* (y - Ax) \cdot w = 0
\]
which shows that \( A^* (y - Ax) = 0 \). (Why?) Therefore, by assumption,
\[
(y - Ax) \cdot y = 0.
\]
Now by (8.2) with \( w = x \),
\[
(y - Ax) \cdot (y - Ax) = (y - Ax) \cdot y - (y - Ax) \cdot Ax = 0
\]
showing that \( y = Ax \). This proves the theorem.

The following corollary is also called the Fredholm alternative.

**Corollary 8.1.10** Let \( A \) be an \( m \times n \) matrix. Then \( A \) is onto if and only if \( A^* \) is one to one.

**Proof:** Suppose first \( A \) is onto. Then by Theorem 8.1.9, it follows that for all \( y \in F^m \), \( y \cdot z = 0 \) whenever \( A^* z = 0 \). Therefore, let \( y = z \) where \( A^* z = 0 \) and conclude that \( z \cdot z = 0 \) whenever \( A^* z = 0 \). If \( A^* x = A^* y \), then \( A^* (x - y) = 0 \) and so \( x - y = 0 \). Thus \( A^* \) is one to one.

Now let \( y \in F^m \) be given. \( y \cdot z = 0 \) whenever \( A^* z = 0 \) because, since \( A^* \) is assumed to be one to one, and 0 is a solution to this equation, it must be the only solution. Therefore, by Theorem 8.1.9 there exists \( x \) such that \( Ax = y \) therefore, \( A \) is onto.

### 8.2 The Dual Basis

That which follows on the dual basis can be extended to \( C^n \) but this will not be done here.

Given a basis, there is something called a dual basis which is very important in applications. A given basis might not be orthonormal so you can’t say \( v_i \cdot v_j = \delta_{ij} \) but you really want to say this. A useful way of getting many of the same advantages is to define something called a dual basis. A dual basis for \( \{v_1, \ldots, v_n\} \) is a set of vectors, \( \{v^1, \ldots, v^n\} \) which has the property that \( v^k \cdot v_j = \delta_{kj} \). Here
\[
\delta_{ij} = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases}
\]
Similarly,
\[
\delta_{ji} = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases}
\]
In this subject it is convenient as well as traditional to keep track of the level on which the index occurs. Thus \( v_i \neq v^i \) ! Of course, sometimes these are the same but not generally.

**Definition 8.2.1** Let \( \{e_i\}_{i=1}^n \) form a basis for \( R^n \). Then \( \{e^i\}_{i=1}^n \) is called the dual basis if
\[
e^i \cdot e_j = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}
\]

**Theorem 8.2.2** If \( \{e_i\}_{i=1}^n \) is a basis then \( \{e^i\}_{i=1}^n \) is also a basis provided (8.3) holds. Furthermore, for each vector, \( v \),
\[
v = (v \cdot e_j) e^j.
\]
Also,
\[
v = (v \cdot e^j) e_j
\]
**Proof:** First we verify that \( \{ e^i \}_{i=1}^n \) is linearly independent. Suppose
\[
0 = v_i e^i. \tag{8.4}
\]
Then taking the dot product of both sides of (8.4) with \( e_j \), yields
\[
0 = v_i e^i \cdot e_j = v_i \delta^i_j = v_j.
\]
Since \( j \) was arbitrary, this shows each \( v_j = 0 \) and so the set is linearly independent as claimed.

It remains to verify \( \{ e^i \}_{i=1}^n \) spans \( \mathbb{R}^n \). Let \( v \in \mathbb{R}^n \) be arbitrary and consider the element in the span of these vectors, \( (v \cdot e_j) e^j \). Then
\[
(v - (v \cdot e_j) e^j) \cdot e_k = v \cdot e_k - (v \cdot e_j) e^j \cdot e_k = v \cdot e_k - (v \cdot e_j) \delta^k_j = 0
\]
and so, since \( \{ e_i \}_{i=1}^n \) is a basis,
\[
(v - v \cdot e_j e^j) \cdot w = 0
\]
for all vectors, \( w \). In particular, this would hold for \( w = (v - v \cdot e_j e^j) \). It follows \( v - v \cdot e_j e^j = 0 \) and this shows \( \{ e^i \}_{i=1}^n \) is a basis.

In the above argument we obtained formulas for the components of a vector \( v \), \( v_i \), with respect to the dual basis, \( (v \cdot e_j) e^j \). Let \( v \) be any vector and let
\[
v = v^j e_j. \tag{8.5}
\]
Then taking the dot product of both sides of (8.5) with \( e^i \), \( v^i = e^i \cdot v \).

Does there exist a dual basis and is it uniquely determined?

**Lemma 8.2.3** Let \( \{ e_i \}_{i=1}^n \) be a basis for \( \mathbb{R}^n \). The matrix, \( G \equiv (g_{ij}) = (e_i \cdot e_j) \) is an invertible matrix. Furthermore \( \det G > 0 \).

**Proof:** Each of these vectors is in \( \mathbb{R}^n \) and so can be written as a column matrix of numbers with respect to the usual basis for \( \mathbb{R}^n \). Now note \( e_i \cdot e_j = e_i^T e_j \). Therefore, the above matrix is nothing more than
\[
G = (g_{ij}) = \begin{pmatrix} e_1^T \\ \vdots \\ e_n^T \end{pmatrix} \begin{pmatrix} e_1 & \cdots & e_n \end{pmatrix} = U^T U
\]
where \( U \) is the matrix which has the \( e_i \) as columns. Therefore,
\[
\det (G) = \det (U^T) \det (U) = \det (U)^2 \geq 0.
\]
Since \( \{ e_i \}_{i=1}^n \) is a basis, it follows the matrix, \( U \) above is one to one. Therefore, the matrix, \( U \) has an inverse and so \( \det (U) \neq 0 \). It follows \( \det (G) > 0 \).

**Definition 8.2.4** The matrix, \( G \) above is called the metric tensor. Its inverse, \( G^{-1} \) is denoted by \( (g^{ij}) \). That is the \( ij \text{th} \) entry of \( G^{-1} \) is denoted as \( g^{ij} \). Thus from the definition of matrix multiplication,
\[
g^{ik} g_{kj} = \delta^i_j.
Theorem 8.2.5 If \( \{ e_i \}_{i=1}^n \) is a basis for \( \mathbb{R}^n \), then there exists a unique dual basis, \( \{ e^i \}_{i=1}^n \) satisfying
\[
e^i \cdot e_i = \delta^i_i.
\]
Furthermore, \( e^i = g^{ij} e_j \).

Proof: \( g^{ij} e_j \cdot e_k = g^{ij} g_{jk} = \delta^i_k \).

This proves the existence of the dual basis. Uniqueness was established earlier.

If \( v \) is any vector, there exist unique scalars, \( v_k \) such that \( v_k e_k = v \). Also, scalars \( v^k \) such that \( v^k e_k = v \). We saw these scalars are given by \( v_k = v \cdot e_k \) and \( v^k = v \cdot e^k \). Do you begin to get the idea on the notation? You sum over indices on different levels.

Definition 8.2.6 If \( v \) is any vector,
\[
v = v_j e^j, \quad v = v^j e_j.
\]
(8.6)
The components of \( v \) which have the index on the top are called the contravariant components of the vector while the components which have the index on the bottom are called the covariant components. In general \( v_i \neq v^j \! \).  

Theorem 8.2.7 The following hold.
\[
g^{ij} e_j = e^i, \quad g_{ij} e^j = e_i, \quad (8.7)
g^{ij} v_j = v^i, \quad g_{ij} v^j = v_i, \quad (8.8)
det (g_{ij}) > 0, \quad det (g^{ij}) > 0. \quad (8.9)
g^{ij} = e^i \cdot e^j, \quad g_{ij} = e_i \cdot e_j \quad (8.10)
\]

Proof: It was shown that \( g^{ij} e_j = e^i \) in Theorem 8.2.5. The second claim of (8.7) follows from Theorem 8.2.2.
\[
e_i = (e_i \cdot e_j) e^j = g_{ij} e^j.
\]
This verifies (8.7). To verify (8.8), use Theorem 8.2.2.
\[
v^i = e^i \cdot v = g^{ij} e_j \cdot v = g^{ij} v_j.
\]
To establish the other formula in (8.8), use Theorem 8.2.2 again.
\[
v_i = e_i \cdot v = g_{ij} e^j \cdot v = g_{ij} v^j.
\]
It was shown that \( det (g_{ij}) > 0 \) already. But \( det (g^{ij}) det (g_{ij}) = 1 \) so this proves the second claim in (8.9). The second of the claims in (8.10) is the way \( g_{ij} \) was defined. It only remains to verify the first equation. Using Theorem 8.2.2,
\[
(e^i \cdot e^k) (e_k \cdot e_j) = (e^i \cdot e^k) (e_k \cdot e_j) = e^i \cdot e_j = \delta^i_j.
\]
Since \( (e^i \cdot e^k) \) acts like the inverse of \( (g_{ij}) \) it follows it is the inverse. This proves the theorem.

The process of writing \( g^{ij} v_j = v^i \) is sometimes called raising the index while the process \( g_{ij} v^j = v_i \) is called lowering the index.
8.2. \textbf{THE DUAL BASIS}  \hspace{1cm} 165

\textbf{Example 8.2.8} Let $\mathbf{e}_1 = (1, 2, 1)^T$, $\mathbf{e}_2 = (0, 1, 1)^T$, and $\mathbf{e}_3 = (3, -1, 1)^T$. Find the dual basis.

As explained above, the metric tensor is

$$ G = \begin{pmatrix} 6 & 3 & 2 \\ 3 & 2 & 0 \\ 2 & 0 & 11 \end{pmatrix} $$

Taking the inverse,

$$ G^{-1} = \begin{pmatrix} \frac{22}{25} & -\frac{33}{25} & -\frac{4}{25} \\ \frac{22}{25} & -\frac{33}{25} & -\frac{4}{25} \\ \frac{22}{25} & -\frac{33}{25} & -\frac{4}{25} \end{pmatrix} $$

Therefore,

$$ \mathbf{e}_1^1 = g^{1j} \mathbf{e}_j = \frac{22}{25} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + -\frac{33}{25} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + -\frac{4}{25} \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{2}{5} \\ \frac{3}{5} \\ \frac{1}{5} \end{pmatrix} $$

$$ \mathbf{e}_2^1 = g^{2j} \mathbf{e}_j = -\frac{33}{25} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + \frac{62}{25} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + \frac{6}{25} \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{3}{5} \\ \frac{3}{5} \\ \frac{1}{5} \end{pmatrix} $$

$$ \mathbf{e}_3^1 = g^{3j} \mathbf{e}_j = -\frac{4}{25} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + \frac{6}{25} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + \frac{3}{25} \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{3}{5} \\ \frac{3}{5} \\ \frac{1}{5} \end{pmatrix} $$

Another way to find the dual basis is as follows. First make the matrix, $M$ which has as columns the given basis. Then multiply on the right by $G^{-1}$. The resulting matrix will have as columns the dual basis. For example, this procedure yields $\mathbf{e}_1^1 = (1, 2, 1)^T$, $\mathbf{e}_2^1 = (0, 1, 1)^T$, and $\mathbf{e}_3^1 = (3, -1, 1)^T$.

This follows from the symmetry of $G^{-1}$ and the definition of matrix multiplication.

\textbf{Example 8.2.9} Let $\mathbf{e}_1 = (1, 2, 0, 1)^T$, $\mathbf{e}_2 = (2, 1, 0, 0)^T$, $\mathbf{e}_3 = (0, 1, 1, 2)^T$, and $\mathbf{e}_4 = (0, 0, 3, 1)^T$. Find the dual basis and metric tensor.

First find the metric tensor. As explained above, this can be done by multiplying the two matrices,

$$ \begin{pmatrix} 1 & 0 & 3 \\ 2 & 1 & -1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 22 & -33 & -4 \\ -14 & 25 & -1 \end{pmatrix} \begin{pmatrix} 12 & 0 & 0 \\ 21 & 11 & 0 \\ 00 & 11 & 2 \\ 00 & 31 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 4 & 4 & 1 \\ 4 & 5 & 1 & 0 \\ 4 & 1 & 6 & 5 \\ 1 & 0 & 5 & 10 \end{pmatrix} $$

Now you invert this matrix to get the inverse of the metric tensor.

$$ \begin{pmatrix} \frac{55}{27} & -\frac{35}{27} & -\frac{5}{3} & \frac{17}{27} \\ -\frac{35}{27} & \frac{28}{27} & \frac{1}{3} & -\frac{10}{27} \\ -\frac{7}{9} & 1 & 3 & -\frac{7}{9} \\ \frac{14}{27} & -\frac{10}{27} & -\frac{2}{3} & \frac{16}{27} \end{pmatrix} $$
Then the dual basis consists of the columns of the matrix,
\[
\begin{pmatrix}
1 & 2 & 0 & 0 \\
2 & 1 & 1 & 0 \\
0 & 0 & 1 & 3 \\
1 & 0 & 2 & 1
\end{pmatrix}
\begin{pmatrix}
\frac{55}{27} & -\frac{35}{27} & -\frac{5}{3} & \frac{17}{27} \\
-\frac{35}{27} & \frac{25}{27} & 1 & -\frac{10}{27} \\
-\frac{5}{3} & \frac{5}{3} & \frac{8}{3} & -\frac{3}{3} \\
\frac{17}{27} & -\frac{10}{27} & -\frac{2}{3} & \frac{10}{27}
\end{pmatrix}
= \begin{pmatrix}
-\frac{5}{9} & \frac{7}{9} & \frac{1}{3} & -\frac{1}{3} \\
\frac{5}{9} & -\frac{5}{9} & -\frac{2}{3} & \frac{2}{3} \\
\frac{7}{3} & -\frac{7}{3} & 1 & -\frac{1}{3}
\end{pmatrix}.
\] (8.11)

Let's check this by multiplying by the matrix which has rows equal to the basis.
\[
\begin{pmatrix}
1 & 2 & 0 & 1 \\
2 & 1 & 0 & 0 \\
0 & 1 & 1 & 2 \\
0 & 0 & 3 & 1
\end{pmatrix}
\begin{pmatrix}
-\frac{5}{9} & \frac{7}{9} & \frac{1}{3} & -\frac{1}{3} \\
\frac{5}{9} & -\frac{5}{9} & -\frac{2}{3} & \frac{2}{3} \\
\frac{7}{3} & -\frac{7}{3} & 1 & -\frac{1}{3}
\end{pmatrix}
= \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}.
\]

Therefore, the columns of (8.11) are the dual basis as hoped.

**Example 8.2.10** In the above example, find the covariant and contravariant components of the vector, \((2, 1, 1) = 2\mathbf{i} + \mathbf{j} + \mathbf{k}.

\[
v_1 = \mathbf{e}_1 \cdot \mathbf{v} = (1, 2, 1) \cdot (2, 1, 1) = 5
\]

\[
v_2 = \mathbf{e}_2 \cdot \mathbf{v} = (0, 1, 1) \cdot (2, 1, 1) = 2
\]

\[
v_3 = \mathbf{e}_3 \cdot \mathbf{v} = (3, -1, 1) \cdot (2, 1, 1) = 6
\]

These are the covariant components. To check whether these work, form
\[
5 \begin{pmatrix}
\frac{2}{3} \\
-\frac{2}{3} \\
\frac{1}{3}
\end{pmatrix} + 2 \begin{pmatrix}
-\frac{1}{3} \\
\frac{1}{3} \\
-\frac{1}{3}
\end{pmatrix} + 6 \begin{pmatrix}
\frac{1}{3} \\
-\frac{1}{3} \\
\frac{1}{3}
\end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}.
\]

Success! Now find the contravariant components. To do this, you can simply raise the index.
\[
\begin{pmatrix}
v^1 \\
v^2 \\
v^3
\end{pmatrix} = \begin{pmatrix}
\frac{22}{27} & -\frac{33}{27} & -\frac{4}{27} \\
-\frac{22}{27} & \frac{33}{27} & \frac{4}{27} \\
-\frac{22}{27} & -\frac{33}{27} & \frac{4}{27}
\end{pmatrix} \begin{pmatrix} 5 \\ 6 \\ 2 \end{pmatrix} = \begin{pmatrix} 4/3 \\ 1/3 \\ -1/3 \end{pmatrix}.
\]

Did it work?
\[
\begin{pmatrix}
\frac{1}{5} \\
\frac{2}{5} \\
\frac{1}{5}
\end{pmatrix} + \begin{pmatrix}
0 \\
1 \\
1
\end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}
\]

Again, success has occurred.

The reason \(G\) is called the metric tensor is contained in the following proposition.
Proposition 8.2.11 Let \( \mathbf{v}, \mathbf{w} \) be two vectors in \( \mathbb{R}^n \) and let \( \{ \mathbf{e}_i \}_{i=1}^n \) be a basis for \( \mathbb{R}^n \). Then
\[
\mathbf{v} \cdot \mathbf{w} = g^{ij} v_i w_j = g^{ij} v_i w_j = g^{ik} v_i w_k.
\]

Proof:
\[
\mathbf{v} \cdot \mathbf{w} = v_i e^i \cdot w_k e^k = e^i \cdot e^k v_i w_k = g^{ik} v_i w_k.
\]
This proves the first equation. The second is similar.

If \( \{ \mathbf{e}_i \}_{i=1}^n \) is the usual orthonormal basis for \( \mathbb{R}^n \), then the metric tensor is just the identity matrix and so you get the usual version of the dot product. The metric tensor allows you to consider the dot product in terms of components of arbitrary bases.

8.3 Exercises

1. The proof of Theorem 8.1.5 concluded with the following observation. If \(-ta + t^2 b \geq 0\) for all \( t \in \mathbb{R} \) and \( b \geq 0 \), then \( a = 0 \). Why is this so?

2. In the proof of Theorem 8.1.9 the following argument was used. If \( \mathbf{x} \cdot \mathbf{w} = 0 \) for all \( \mathbf{w} \in \mathbb{R}^n \), then \( \mathbf{x} = \mathbf{0} \). Why is this so?

3. Suppose \( L : \mathbb{R}^n \rightarrow \mathbb{R}^m \) is a linear transformation. Show the following are equivalent.
   
   (a) \( L \mathbf{x} = \mathbf{0} \) implies \( \mathbf{x} = \mathbf{0} \).

   (b) \( L \) is one to one.

4. Using Corollary 8.1.10 and Problem 3, show that an \( m \times n \) matrix is onto if and only if its transpose is one to one.

5. Suppose \( A \) is a \( 3 \times 2 \) matrix. Is it possible that \( A^T \) is one to one? What does this say about \( A \) being onto? Prove your answer.

6. Explain why there always exists a solution to the equation, \( A^T \mathbf{y} = A^T \mathbf{A} \mathbf{x} \) and also explain why this is called a least squares solution to the equation, \( \mathbf{y} = \mathbf{A} \mathbf{x} \).

7. Referring to Problem 6, find the least squares solution to the following system.
   
   \[
   \begin{align*}
   x + 2y &= 1 \\
   2x + 3y &= 2 \\
   3x + 5y &= 4
   \end{align*}
   \]

8. You are doing experiments and have obtained the ordered pairs, \((0, 1), (1, 2), (2, 3.5), \) and \((3, 4)\). Find \( m \) and \( b \) such that \( y = mx + b \) approximates these four points as well as possible. Now do the same thing for \( y = ax^2 + bx + c \), finding \( a, b, \) and \( c \) to give the best approximation.

9. Suppose you have several ordered triples, \((x_i, y_i, z_i)\). Describe how to find a polynomial,
   
   \[ z = a + bx + cy + dx + ey^2 \]

   for example giving the best fit to the given ordered triples. Is there any reason you have to use a polynomial? Would similar approaches work for other combinations of functions just as well?

10. Using the Gram Schmidt process, find an orthonormal basis for the span of the vectors, \((1, 2, 1), (2, -1, 3), \) and \((1, 0, 0)\).
11. Using the Gram Schmidt process, find an orthonormal basis for the span of the vectors,
\((1, 2, 1, 0), (2, -1, 3, 1), \text{ and } (1, 0, 0, 1)\).

12. The set, \(V \equiv \{(x, y, z) : 2x + 3y - z = 0\}\) is a subspace of \(\mathbb{R}^3\). Find an orthonormal basis for this subspace.

13. The two level surfaces, \(2x + 3y - z + w = 0\) and \(3x - y + z + 2w = 0\) intersect in a subspace of \(\mathbb{R}^3\), find a basis for this subspace. Next find an orthonormal basis for this subspace.

14. Let \(\{v_1, \ldots, v_m\}\) be a linearly independent set of vectors. Let \(u_1 = v_1\) and if \(u_1, \ldots, u_k\) have been chosen for \(k < m\), define
\[
 u_{k+1} \equiv v_{k+1} - \sum_{j=1}^{k} \frac{v_{k+1} \cdot u_j}{|u_j|^2} u_j.
\]

Show that each \(u_k\) is non zero, \(u_k \cdot u_l = 0\) if \(k \neq l\), and for each \(k \leq m\)
\[
\text{span } (v_1, \ldots, v_k) = \text{span } (u_1, \ldots, u_k).
\]

15. Let \(e_1 = i + j, e_2 = i - j, e_3 = j + k\). Find \(e^1, e^2, e^3, (g_{ij}), (g^{ij})\). If \(v = i + 2j + k\), find \(v^i\) and \(v_j\), the contravariant and covariant components of the vector.

16. Let \(e^1 = 2i + j, e^2 = i - 2j, e^3 = k\). Find \(e_1, e_2, e_3, (g_{ij}), (g^{ij})\). If \(v = 2i - 2j + k\), find \(v^i\) and \(v_j\), the contravariant and covariant components of the vector.

17. Suppose \(e_1, e_2, e_3\) have the property that \(e_i \cdot e_j = 0\) whenever \(i \neq j\). Show that then the metric tensor is a diagonal matrix.

18. Suppose \(e_1, e_2, e_3\) have the property that \(e_i \cdot e_j = 0\) whenever \(i \neq j\). Show the same is true of the dual basis and that in fact, \(e^i\) is a multiple of \(e_i\).

19. Let \(v = v_i e^i\) and let \(w = w^j e_j\). Show that \(v \cdot w = v_i w^i\).

20. Show if \(\{e_i\}_{i=1}^3\) is a basis in \(\mathbb{R}^3\)
\[
e^1 = \frac{e_2 \times e_3}{e_2 \times e_3 \cdot e_1}, \quad e^2 = \frac{e_1 \times e_2}{e_1 \times e_3 \cdot e_2}, \quad e^3 = \frac{e_1 \times e_2}{e_1 \times e_2 \cdot e_3}.
\]

21. Let \(\{e_i\}_{i=1}^n\) be a basis and define
\[
e_i^* \equiv \frac{e_i}{|e_i|}, \quad e^* i \equiv e^i |e_i|.
\]

Show \(e^* i \cdot e_j^* = \delta^i_j\).

22. If \(v\) is a vector, \(v_i^*\) and \(v^* i\), are defined by
\[
v \equiv v_i^* e^* i \equiv v^* i e_i^*.
\]

These are called the physical components of \(v\). Show
\[
v_i^* = \frac{v_i}{|e_i|}, \quad v^* i = \frac{v_i}{|e_i|} \quad (\text{No summation on } i).
\]
Linear Transformations

9.0.1 Outcomes

1. Define linear transformation. Interpret a matrix as a linear transformation.
2. Find a matrix that represents a linear transformation given by a geometric description.
3. Write the solution space of a homogeneous system as the span of a set of basis vectors.
   Determine the dimension of the solution space.
4. Relate the solutions of a non-homogeneous system to the solutions of a homogeneous system.

9.1 Linear Transformations

An \( m \times n \) matrix can be used to transform vectors in \( \mathbb{F}^n \) to vectors in \( \mathbb{F}^m \) through the use of matrix multiplication.

**Example 9.1.1** Consider the matrix, \[
\begin{pmatrix}
1 & 2 & 0 \\
2 & 1 & 0
\end{pmatrix}
\]. Think of it as a function which takes vectors in \( \mathbb{F}^3 \) and makes them in to vectors in \( \mathbb{F}^2 \) as follows. For \[
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
\] a vector in \( \mathbb{F}^3 \), multiply on the left by the given matrix to obtain the vector in \( \mathbb{F}^2 \). Here are some numerical examples.

\[
\begin{pmatrix}
1 & 2 & 0 \\
2 & 1 & 0
\end{pmatrix}
\begin{pmatrix}
1 \\
2
\end{pmatrix} = \begin{pmatrix}
5 \\
4
\end{pmatrix},
\begin{pmatrix}
1 & 2 & 0 \\
2 & 1 & 0
\end{pmatrix}
\begin{pmatrix}
1 \\
2
\end{pmatrix} = \begin{pmatrix}
-2 \\
3
\end{pmatrix},
\begin{pmatrix}
1 & 2 & 0 \\
2 & 1 & 0
\end{pmatrix}
\begin{pmatrix}
10 \\
5
\end{pmatrix} = \begin{pmatrix}
20 \\
25
\end{pmatrix},
\begin{pmatrix}
1 & 2 & 0 \\
2 & 1 & 0
\end{pmatrix}
\begin{pmatrix}
0 \\
7
\end{pmatrix} = \begin{pmatrix}
0 \\
7
\end{pmatrix},\begin{pmatrix}
14 \\
7
\end{pmatrix}.
\]

More generally,

\[
\begin{pmatrix}
1 & 2 & 0 \\
2 & 1 & 0
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix} = \begin{pmatrix}
x + 2y \\
2x + y
\end{pmatrix}.
\]

The idea is to define a function which takes vectors in \( \mathbb{F}^3 \) and delivers new vectors in \( \mathbb{F}^2 \).

This is an example of something called a linear transformation.
Definition 9.1.2 Let $X$ and $Y$ be vector spaces and let $T : X \to Y$ be a function. Thus for each $x \in X, Tx \in Y$. Then $T$ is a linear transformation if whenever $\alpha, \beta$ are scalars and $x_1$ and $x_2$ are vectors in $X$,

$$T(\alpha x_1 + \beta x_2) = \alpha_1 T x_1 + \beta T x_2.$$  

In words, linear transformations distribute across $+$ and allow you to factor out scalars.

At this point, recall the properties of matrix multiplication. The pertinent property is (3.14) on Page 47. Recall it states that for $a$ and $b$ scalars,

$$A(ab + bc) = aAB + bAC.$$  

In particular, for $A$ an $m \times n$ matrix and $B$ and $C, n \times 1$ matrices (column vectors) the above formula holds which is nothing more than the statement that matrix multiplication gives an example of a linear transformation.

Definition 9.1.3 A linear transformation is called one to one (often written as $1 \rightarrow 1$) if it never takes two different vectors to the same vector. Thus $T$ is one to one if whenever $x \neq y$

$$Tx \neq Ty.$$  

Equivalently, if $T(x) = T(y)$, then $x = y$.

In the case that a linear transformation comes from matrix multiplication, it is common usage to refer to the matrix as a one to one matrix when the linear transformation it determines is one to one.

Definition 9.1.4 A linear transformation mapping $X$ to $Y$ is called onto if whenever $y \in Y$ there exists $x \in X$ such that $T(x) = y$.

Thus $T$ is onto if everything in $Y$ gets hit. In the case that a linear transformation comes from matrix multiplication, it is common to refer to the matrix as onto when the linear transformation it determines is onto. Also it is common usage to write $TX$, $T(X)$, or $\text{Im}(T)$ as the set of vectors of $Y$ which are of the form $Tx$ for some $x \in X$. In the case that $T$ is obtained from multiplication by an $m \times n$ matrix, $A$, it is standard to simply write $A(F^n)$ or $\text{Im}(A)$ to denote those vectors in $F^m$ which are obtained in the form $Ax$ for some $x \in F^n$.

9.2 Constructing The Matrix Of A Linear Transformation

It turns out that if $T$ is any linear transformation which maps $F^n$ to $F^m$, there is always an $m \times n$ matrix, $A$ with the property that

$$Ax = Tx$$  

(9.1)

for all $x \in F^n$. Here is why. Suppose $T : F^n \to F^m$ is a linear transformation and you want to find the matrix defined by this linear transformation as described in (9.1). Then if $x \in F^n$ it follows

$$x = \sum_{i=1}^{n} x_i e_i.$$  

In words, linear transformations distribute across $+$ and allow you to factor out scalars.
where $e_i$ is the vector which has zeros in every slot but the $i^{th}$ and a 1 in this slot. Then since $T$ is linear,

$$T\mathbf{x} = \sum_{i=1}^{n} x_i T(e_i)$$

$$= \begin{pmatrix} T(e_1) & \cdots & T(e_n) \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

$$= A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

and so you see that the matrix desired is obtained from letting the $i^{th}$ column equal $T(e_i)$.

We state this as the following theorem.

**Theorem 9.2.1** Let $T$ be a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$. Then the matrix, $A$ satisfying (9.1) is given by

$$\begin{pmatrix} T(e_1) & \cdots & T(e_n) \end{pmatrix}$$

where $Te_i$ is the $i^{th}$ column of $A$.

### 9.2.1 Rotations Of $\mathbb{R}^2$

Sometimes you need to find a matrix which represents a given linear transformation which is described in geometrical terms. The idea is to produce a matrix which you can multiply a vector by to get the same thing as some geometrical description. A good example of this is the problem of rotation of vectors.

**Example 9.2.2** Determine the matrix which represents the linear transformation defined by rotating every vector through an angle of $\theta$.

Let $e_1 \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $e_2 \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. These identify the geometric vectors which point along the positive $x$ axis and positive $y$ axis as shown.

```
 e2
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
  |
From the above, you only need to find $T e_1$ and $T e_2$, the first being the first column of the desired matrix, $A$ and the second being the second column. From drawing a picture and doing a little geometry, you see that $T e_1 = \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix}$, $T e_2 = \begin{pmatrix} -\sin \theta \\ \cos \theta \end{pmatrix}$.

Therefore, from Theorem 9.2.1,

$$A = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

**Example 9.2.3** Find the matrix of the linear transformation which is obtained by first rotating all vectors through an angle of $\phi$ and then through an angle $\theta$. Thus you want the linear transformation which rotates all angles through an angle of $\theta + \phi$.

Let $T_{\theta+\phi}$ denote the linear transformation which rotates every vector through an angle of $\theta + \phi$. Then to get $T_{\theta+\phi}$, you could first do $T_{\phi}$ and then do $T_{\theta}$ where $T_{\phi}$ is the linear transformation which rotates through an angle of $\phi$ and $T_{\theta}$ is the linear transformation which rotates through an angle of $\theta$. Denoting the corresponding matrices by $A_{\theta+\phi}$, $A_{\phi}$, and $A_{\theta}$, you must have for every $x$

$$A_{\theta+\phi}x = T_{\theta+\phi}x = T_{\theta}T_{\phi}x = A_{\theta}A_{\phi}x.$$ 

Consequently, you must have

$$A_{\theta+\phi} = \begin{pmatrix} \cos (\theta + \phi) & -\sin (\theta + \phi) \\ \sin (\theta + \phi) & \cos (\theta + \phi) \end{pmatrix} = A_{\theta}A_{\phi}$$

$$= \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix}.$$ 

You know how to multiply matrices. Do so to the pair on the right. This yields

$$\begin{pmatrix} \cos \theta + \phi & -\sin \theta + \phi \\ \sin \theta + \phi & \cos \theta + \phi \end{pmatrix} = \begin{pmatrix} \cos \theta \cos \phi - \sin \theta \sin \phi & -\cos \theta \sin \phi - \sin \theta \cos \phi \\ \sin \theta \cos \phi + \cos \theta \sin \phi & \cos \theta \cos \phi - \sin \theta \sin \phi \end{pmatrix}.$$ 

Don’t these look familiar? They are the usual trig. identities for the sum of two angles derived here using linear algebra concepts.

You do not have to stop with two dimensions. You can consider rotations and other geometric concepts in any number of dimensions. This is one of the major advantages of linear algebra. You can break down a difficult geometrical procedure into small steps, each corresponding to multiplication by an appropriate matrix. Then by multiplying the matrices, you can obtain a single matrix which can give you numerical information on the results of applying the given sequence of simple procedures. That which you could never visualize can still be understood to the extent of finding exact numerical answers. Another example follows.

### 9.2.2 Projections

In Physics it is important to consider the work done by a force field on an object. This involves the concept of projection onto a vector. Suppose you want to find the projection of a vector, $v$, onto the given vector, $u$, denoted by $\text{proj}_u(v)$ This is done using the dot product as follows.

$$\text{proj}_u(v) = \left( \frac{v \cdot u}{u \cdot u} \right) u$$
Because of properties of the dot product, the map \( v \rightarrow \text{proj}_u(v) \) is linear,
\[
\text{proj}_u(\alpha v + \beta w) = \left( \frac{\alpha v + \beta w \cdot u}{u \cdot u} \right) u = \alpha \left( \frac{v \cdot u}{u \cdot u} \right) u + \beta \left( \frac{w \cdot u}{u \cdot u} \right) u
\]
\[= \alpha \text{proj}_u(v) + \beta \text{proj}_u(w).
\]

**Example 9.2.4** Let the projection map be defined above and let \( u = (1, 2, 3)^T \). Does this linear transformation come from multiplication by a matrix? If so, what is the matrix?

You can find this matrix in the same way as in the previous example. Let \( e_i \) denote the vector in \( \mathbb{R}^n \) which has a 1 in the \( i \)th position and a zero everywhere else. Thus a typical vector, \( x = (x_1, \cdots, x_n)^T \) can be written in a unique way as
\[
x = \sum_{j=1}^{n} x_j e_j.
\]

From the way you multiply a matrix by a vector, it follows that \( \text{proj}_u(e_i) \) gives the \( i \)th column of the desired matrix. Therefore, it is only necessary to find
\[
\text{proj}_u(e_i) \equiv \left( \frac{e_i \cdot u}{u \cdot u} \right) u
\]

For the given vector in the example, this implies the columns of the desired matrix are
\[
\frac{1}{14} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \frac{2}{14} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \frac{3}{14} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.
\]

Hence the matrix is
\[
\frac{1}{14} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{pmatrix}.
\]

**9.2.3 Matrices Which Are One To One Or Onto**

**Lemma 9.2.5** Let \( A \) be an \( m \times n \) matrix. Then \( A(\mathbb{R}^n) = \text{span}(a_1, \cdots, a_n) \) where \( a_1, \cdots, a_n \) denote the columns of \( A \). In fact, for \( x = (x_1, \cdots, x_n)^T \),
\[
Ax = \sum_{k=1}^{n} x_k a_k.
\]

**Proof:** This follows from the definition of matrix multiplication in Definition 3.1.9 on Page 42.

The following is a theorem of major significance. First here is an interesting observation.

**Observation 9.2.6** Let \( A \) be an \( m \times n \) matrix. Then \( A \) is one to one if and only if \( Ax = 0 \) implies \( x = 0 \).

Here is why: \( A0 = A(0+0) = A0 + A0 \) and so \( A0 = 0 \).

Now suppose \( A \) is one to one and \( Ax = 0 \). Then since \( A0 = 0 \), it follows \( x = 0 \). Thus if \( A \) is one to one and \( Ax = 0 \), then \( x = 0 \).

Next suppose the condition that \( Ax = 0 \) implies \( x = 0 \) is valid. Then if \( Ax = Ay \), then \( A(x-y) = 0 \) and so from the condition, \( x - y = 0 \) so that \( x = y \). Thus \( A \) is one to one.
Theorem 9.2.7 Suppose \( A \) is an \( n \times n \) matrix. Then \( A \) is one to one if and only if \( A \) is onto. Also, if \( B \) is an \( n \times n \) matrix and \( AB = I \), then it follows \( BA = I \).

**Proof:** First suppose \( A \) is one to one. Consider the vectors, \( \{Ae_1, \ldots, Ae_n\} \) where \( e_k \) is the column vector which is all zeros except for a 1 in the \( k^{th} \) position. This set of vectors is linearly independent because if

\[
\sum_{k=1}^{n} c_k Ae_k = 0,
\]

then since \( A \) is linear,

\[
A \left( \sum_{k=1}^{n} c_k e_k \right) = 0
\]

and since \( A \) is one to one, it follows

\[
\sum_{k=1}^{n} c_k e_k = 0
\]

which implies each \( c_k = 0 \). Therefore, \( \{Ae_1, \ldots, Ae_n\} \) must be a basis for \( \mathbb{F}^n \) because if not there would exist a vector, \( y \notin \text{span} \{Ae_1, \ldots, Ae_n\} \) and then by Lemma 5.1.41, \( \{Ae_1, \ldots, Ae_n, y\} \) would be an independent set of vectors having \( n+1 \) vectors in it, contrary to the exchange theorem. It follows that for \( y \in \mathbb{F}^n \) there exist constants, \( c_i \) such that

\[
y = \sum_{k=1}^{n} c_k Ae_k = A \left( \sum_{k=1}^{n} c_k e_k \right)
\]

showing that, since \( y \) was arbitrary, \( A \) is onto.

Next suppose \( A \) is onto. By Lemma 9.2.5, this means the span of the columns of \( A \) equals \( \mathbb{F}^n \). If these columns are not linearly independent, then by Lemma 5.1.28 on Page 98, one of the columns is a linear combination of the others and so the span of the columns of \( A \) equals the span of the \( n-1 \) other columns. This violates the exchange theorem because \( \{e_1, \ldots, e_n\} \) would be a linearly independent set of vectors contained in the span of only \( n-1 \) vectors. Therefore, the columns of \( A \) must be independent and by Lemma 9.2.5 this is equivalent to saying that \( Ax = 0 \) if and only if \( x = 0 \). This implies \( A \) is one to one because if \( Ax = Ay \), then \( A(x-y) = 0 \) and so \( x-y = 0 \).

Now suppose \( AB = I \). Why is \( BA = I \)? Since \( AB = I \) it follows \( B \) is one to one since otherwise, there would exist, \( x \neq 0 \) such that \( Bx = 0 \) and then \( ABx = A0 = 0 \neq Ix \). Therefore, from what was just shown, \( B \) is also onto. In addition to this, \( A \) must be one to one because if \( Ay = 0 \), then \( y = Bx \) for some \( x \) and then \( x = ABx = Ay = 0 \) showing \( y = 0 \). Now from what is given to be so, it follows \( (AB)A = A \) and so using the associative law for matrix multiplication,

\[
A(BA) - A = A(BA - I) = 0.
\]

But this means \( (BA - I)x = 0 \) for all \( x \) since otherwise, \( A \) would not be one to one. Hence \( BA = I \) as claimed. This proves the theorem.

This theorem shows that if an \( n \times n \) matrix, \( B \) acts like an inverse when multiplied on one side of \( A \) it follows that \( B = A^{-1} \) and it will act like an inverse on both sides of \( A \).

The conclusion of this theorem pertains to square matrices only. For example, let

\[
A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & -1 \end{pmatrix}
\]

(9.2)
Then
\[ BA = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \]
buts\[ AB = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & -1 \\ 1 & 0 & 0 \end{pmatrix}. \]

There is also an important characterization in terms of determinants.

**Theorem 9.2.8** Let \( A \) be an \( n \times n \) matrix and let \( T_A \) denote the linear transformation determined by \( A \). Then the following are equivalent.

1. \( T_A \) is one to one.
2. \( T_A \) is onto.
3. \( \det(A) \neq 0 \).

### 9.2.4 The General Solution Of A Linear System

Recall the following definition which was discussed above.

**Definition 9.2.9** \( T \) is a linear transformation if whenever \( x, y \) are vectors and \( a, b \) scalars,
\[ T(ax + by) = aTx + bTy. \]
Thus linear transformations distribute across addition and pass scalars to the outside. A linear system is one which is of the form
\[ Tx = b. \]
If \( Tx_p = b \), then \( x_p \) is called a particular solution to the linear system.

For example, if \( A \) is an \( m \times n \) matrix and \( T_A \) is determined by
\[ T_A(x) = Ax, \]
then from the properties of matrix multiplication, \( T_A \) is a linear transformation. In this setting, we will usually write \( A \) for the linear transformation as well as the matrix. There are many other examples of linear transformations other than this. In differential equations, you will encounter linear transformations which act on functions to give new functions. In this case, the functions are considered as vectors.

**Definition 9.2.10** Let \( T \) be a linear transformation. Define
\[ \ker(T) \equiv \{ x : Tx = 0 \}. \]
Thus \( \ker(T) \) consists of the set of all vectors which \( T \) sends to \( 0 \).

The above definition states that \( \ker(T) \) is the set of solutions to the equation,
\[ Tx = 0. \]
In the case where \( T \) is really a matrix, you have been solving such equations for quite some time. However, sometimes linear transformations act on vectors which are not in \( \mathbb{F}^n \).
**Example 9.2.11** Let $\frac{df}{dx}$ denote the linear transformation defined on $X$, the functions which are defined on $\mathbb{R}$ and have a continuous derivative. Find $\ker \left( \frac{df}{dx} \right)$.

The example asks for functions, $f$ which the property that $\frac{df}{dx} = 0$. As you know from calculus, these functions are the constant functions. Thus $\ker \left( \frac{df}{dx} \right) = \text{constant functions}$.

When $T$ is a linear transformation, systems of the form $Tx = 0$ are called **homogeneous systems**. Thus the solution to the homogeneous system is known as $\ker(T)$.

Systems of the form $Tx = b$ where $b \neq 0$ are called **nonhomogeneous systems**. It turns out there is a very interesting and important relation between the solutions to the homogeneous systems and the solutions to the nonhomogeneous systems.

**Theorem 9.2.12** Suppose $x_p$ is a solution to the linear system, $Tx = b$.

Then if $y$ is any other solution to the linear system, there exists $x \in \ker(T)$ such that

$$y = x_p + x.$$ 

**Proof:** Consider $y - x_p \equiv y + (-1)x_p$. Then $T(y - x_p) = Ty - Tx_p = b - b = 0$. Let $x \equiv y - x_p$. This proves the theorem.

Sometimes people remember the above theorem in the following form. The solutions to the nonhomogeneous system, $Tx = b$ are given by $x_p + \ker(T)$ where $x_p$ is a particular solution to $Tx = b$.

We have been vague about what $T$ is and what $x$ is on purpose. This theorem is completely algebraic in nature and will work whenever you have linear transformations. In particular, it will be important in differential equations. For now, here is a familiar example.

**Example 9.2.13** Let

$$A = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 2 & 1 & 1 & 2 \\ 4 & 5 & 7 & 2 \end{pmatrix}$$

Find $\ker(A)$.

This asks you to find $\{x : Ax = 0\}$. In other words you are asked to solve the system, $Ax = 0$. Let $x = (x, y, z, w)^T$. Then this amounts to solving

$$\begin{pmatrix} 1 & 2 & 3 & 0 \\ 2 & 1 & 1 & 2 \\ 4 & 5 & 7 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

This is the linear system

$$x + 2y + 3z = 0$$
$$2x + y + z + 2w = 0$$
$$4x + 5y + 7z + 2w = 0$$

and you know how to solve this using row operations, (Gauss Elimination). Set up the augmented matrix,

$$\begin{pmatrix} 1 & 2 & 3 & 0 & | & 0 \\ 2 & 1 & 1 & 2 & | & 0 \\ 4 & 5 & 7 & 2 & | & 0 \end{pmatrix}$$
Then row reduce to obtain a reduced echelon form,

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

This yields \( x = \frac{1}{3} z - \frac{4}{3} w \) and \( y = \frac{2}{3} w - \frac{5}{3} z \). Thus \( \ker(A) \) consists of vectors of the form,

\[
\begin{pmatrix}
\frac{1}{3} z - \frac{4}{3} w \\
\frac{2}{3} w - \frac{5}{3} z \\
z \\
w
\end{pmatrix} = z
\begin{pmatrix}
\frac{1}{3} \\
\frac{2}{3} \\
1 \\
0
\end{pmatrix} + w
\begin{pmatrix}
\frac{4}{3} \\
\frac{5}{3} \\
0 \\
1
\end{pmatrix}.
\]

Example 9.2.14 The general solution of a linear system of equations is just the set of all solutions. Find the general solution to the linear system,

\[
\begin{pmatrix}
1 & 2 & 3 & 0 \\
2 & 1 & 1 & 2 \\
4 & 5 & 7 & 2
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z \\
w
\end{pmatrix} =
\begin{pmatrix}
9 \\
7 \\
25
\end{pmatrix}
\]

given that \( \begin{pmatrix} 1 & 1 & 2 & 1 \end{pmatrix}^T = \begin{pmatrix} x & y & z & w \end{pmatrix}^T \) is one solution.

Note the matrix on the left is the same as the matrix in Example 9.2.13. Therefore, from Theorem 9.2.12, you will obtain all solutions to the above linear system in the form

\[
z
\begin{pmatrix}
\frac{1}{3} \\
\frac{2}{3} \\
1 \\
0
\end{pmatrix} + w
\begin{pmatrix}
\frac{4}{3} \\
\frac{5}{3} \\
0 \\
1
\end{pmatrix} + \begin{pmatrix}
1 \\
1 \\
2 \\
1
\end{pmatrix}.
\]

9.3 Exercises

1. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( \pi/3 \).
2. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( \pi/4 \).
3. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( -\pi/3 \).
4. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( 2\pi/3 \).
5. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( \pi/12 \). **Hint:** Note that \( \pi/12 = \pi/3 - \pi/4 \).
6. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( 2\pi/3 \) and then reflects across the \( x \) axis.
7. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( \pi/3 \) and then reflects across the \( x \) axis.

8. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( \pi/4 \) and then reflects across the \( x \) axis.

9. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( \pi/6 \) and then reflects across the \( x \) axis followed by a reflection across the \( y \) axis.

10. Find the matrix for the linear transformation which reflects every vector in \( \mathbb{R}^2 \) across the \( x \) axis and then rotates every vector through an angle of \( \pi/4 \).

11. Find the matrix for the linear transformation which reflects every vector in \( \mathbb{R}^2 \) across the \( y \) axis and then rotates every vector through an angle of \( \pi/4 \).

12. Find the matrix for the linear transformation which reflects every vector in \( \mathbb{R}^2 \) across the \( x \) axis and then rotates every vector through an angle of \( \pi/6 \).

13. Find the matrix for the linear transformation which reflects every vector in \( \mathbb{R}^2 \) across the \( y \) axis and then rotates every vector through an angle of \( \pi/6 \).

14. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( \frac{5\pi}{12} \). \text{Hint: Note that } \frac{5\pi}{12} = \frac{2\pi}{3} - \frac{\pi}{4}.

15. Find the matrix for \( \text{proj}_u (v) \) where \( u = (1, -2, 3)^T \).

16. Find the matrix for \( \text{proj}_u (v) \) where \( u = (1, 5, 3)^T \).

17. Find the matrix for \( \text{proj}_u (v) \) where \( u = (1, 0, 3)^T \).

18. Show that the function \( T_u \) defined by \( T_u (v) \equiv v - \text{proj}_u (v) \) is also a linear transformation.

19. If \( u = (1, 2, 3)^T \), as in Example 9.2.4 and \( T_u \) is given in the above problem, find the matrix, \( A_u \) which satisfies \( A_u x = T(x) \).

20. If \( A, B, \) and \( C \) are each \( n \times n \) matrices and \( ABC \) is invertible, why are each of \( A, B, \) and \( C \) invertible.

21. Show that \( (ABC)^{-1} = C^{-1}B^{-1}A^{-1} \) by doing the computation \( ABC \left( C^{-1}B^{-1}A^{-1} \right) \).

22. If \( A \) is invertible, show \( (A^T)^{-1} = (A^{-1})^T \).

23. If \( A \) is invertible, show \( (A^2)^{-1} = (A^{-1})^2 \).

24. If \( A \) is invertible, show \( (A^{-1})^{-1} = A \).

25. 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.

26. Explain why \( Ax = 0 \) always has a solution.

27. Suppose \( \det (A - \lambda I) = 0 \). Show using Theorem 9.2.8 there exists \( x \neq 0 \) such that \( (A - \lambda I)x = 0 \).
28. Let $A$ be an $n \times n$ matrix and let $x$ be a nonzero vector such that $Ax = \lambda x$ for some scalar, $\lambda$. When this occurs, the vector, $x$ is called an **eigenvector** and the scalar, $\lambda$ is called an **eigenvalue**. It turns out that not every number is an eigenvalue. Only certain ones are. Why? **Hint:** Show that if $Ax = \lambda x$, then $(A - \lambda I)x = 0$. Explain why this shows that $(A - \lambda I)$ is not one to one and not onto. Now use Theorem 9.2.8 to argue $\det (A - \lambda I) = 0$. What sort of equation is this? How many solutions does it have?

29. 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 $Ax = 0$.

30. Find $\ker(A)$ for

$$A = \begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 2 & 1 & 1 & 2 \\ 1 & 4 & 4 & 3 & 3 \\ 0 & 2 & 1 & 1 & 2 \end{pmatrix}.$$  

Recall $\ker(A)$ is just the set of solutions to $Ax = 0$.

31. Suppose $Ax = b$ has a solution. Explain why the solution is unique precisely when $Ax = 0$ has only the trivial (zero) solution.

32. Using Problem 30, find the general solution to the following linear system.

$$\begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 2 & 1 & 1 & 2 \\ 1 & 4 & 4 & 3 & 3 \\ 0 & 2 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 11 \\ 7 \\ 18 \\ 7 \end{pmatrix}.$$  

33. Using Problem 30, find the general solution to the following linear system.

$$\begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 2 & 1 & 1 & 2 \\ 1 & 4 & 4 & 3 & 3 \\ 0 & 2 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 6 \\ 7 \\ 13 \\ 7 \end{pmatrix}.$$  

34. Show that if $A$ is an $m \times n$ matrix, then $\ker(A)$ is a subspace.

35. Sometimes it is required to consider rotations in three dimensions. An example is the Euler angles in the mechanics of a rotating body. Describe how you could use the concepts of matrix multiplication to systematically study such rotations.

36. Verify the linear transformation determined by the matrix of (9.2) maps $\mathbb{R}^3$ onto $\mathbb{R}^2$ but the linear transformation determined by this matrix is not one to one.
10.0.1 Outcomes

1. Describe the eigenvalue problem geometrically and algebraically.
2. Evaluate the spectrum and eigenvectors for a square matrix.
3. Use the determinant of the Grammian matrix to compute volumes of \( k \) dimensional parallelepipeds.
4. Recall and use block multiplication.
5. Read and understand the proof of Schur’s theorem.

10.1 Eigenvalues And Eigenvectors Of A Matrix

Spectral Theory refers to the study of eigenvalues and eigenvectors of a matrix. It is of fundamental importance in many areas. Row operations will no longer be such a useful tool in this subject.

10.1.1 Definition Of Eigenvectors And Eigenvalues

In this section, \( F = \mathbb{C} \).

To illustrate the idea behind what will be discussed, consider the following example.

Example 10.1.1 Here is a matrix.

\[
\begin{pmatrix}
0 & 5 & -10 \\
0 & 22 & 16 \\
0 & -9 & -2
\end{pmatrix}.
\]

Multiply this matrix by the vector

\[
\begin{pmatrix}
-5 \\
-4 \\
3
\end{pmatrix}
\]

and see what happens. Then multiply it by

\[
\begin{pmatrix}
1 \\
0 \\
0
\end{pmatrix}
\]

and see what happens. Does this matrix act this way for some other vector?
First
\[
\begin{pmatrix}
0 & 5 & -10 \\
0 & 22 & 16 \\
0 & -9 & -2
\end{pmatrix}
\begin{pmatrix}
-5 \\
-4 \\
3
\end{pmatrix}
= \begin{pmatrix}
-50 \\
-40 \\
30
\end{pmatrix}
= 10 \begin{pmatrix}
-5 \\
-4 \\
3
\end{pmatrix}.
\]

Next
\[
\begin{pmatrix}
0 & 5 & -10 \\
0 & 22 & 16 \\
0 & -9 & -2
\end{pmatrix}
\begin{pmatrix}
1 \\
0 \\
0
\end{pmatrix}
= \begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}
= 0 \begin{pmatrix}
1 \\
1 \\
1
\end{pmatrix}.
\]

When you multiply the first vector by the given matrix, it stretched the vector, multiplying it by 10. When you multiplied the matrix by the second vector it sent it to the zero vector. Now consider
\[
\begin{pmatrix}
0 & 5 & -10 \\
0 & 22 & 16 \\
0 & -9 & -2
\end{pmatrix}
\begin{pmatrix}
1 \\
1 \\
1
\end{pmatrix}
= \begin{pmatrix}
-5 \\
38 \\
-11
\end{pmatrix}.
\]

In this case, multiplication by the matrix did not result in merely multiplying the vector by a number.

In the above example, the first two vectors were called eigenvectors and the numbers, 10 and 0 are called eigenvalues. Not every number is an eigenvalue and not every vector is an eigenvector.

**Definition 10.1.2** Let \( M \) be an \( n \times n \) matrix and let \( x \in \mathbb{C}^n \) be a nonzero vector for which
\[
Mx = \lambda x
\]
for some scalar, \( \lambda \). Then \( x \) is called an eigenvector and \( \lambda \) is called an eigenvalue (characteristic value) of the matrix, \( M \).

**Note:** Eigenvectors are never equal to zero!

The set of all eigenvalues of an \( n \times n \) matrix, \( M \), is denoted by \( \sigma (M) \) and is referred to as the spectrum of \( M \).

The eigenvectors of a matrix \( M \) are those vectors, \( x \) for which multiplication by \( M \) results in a vector in the same direction or opposite direction to \( x \). Since the zero vector, \( 0 \) has no direction this would make no sense for the zero vector. As noted above, \( 0 \) is never allowed to be an eigenvector. How can eigenvectors be identified? Suppose \( x \) satisfies (10.1). Then
\[
(M - \lambda I)x = 0
\]
for some \( x \neq 0 \). (Equivalently, you could write \((\lambda I - M)x = 0\).) Sometimes we will use
\[
(\lambda I - M)x = 0
\]
and sometimes
\[
(M - \lambda I)x = 0.
\]
It makes absolutely no difference and you should use whichever you like better. Therefore, the matrix \( M - \lambda I \) cannot have an inverse because if it did, the equation could be solved,
\[
x = \left( (M - \lambda I)^{-1} (M - \lambda I) \right) x = (M - \lambda I)^{-1} \left( (M - \lambda I)x \right) = (M - \lambda I)^{-1} 0 = 0,
\]

**Note:** Eigenvectors are never equal to zero!
and this would require $x = 0$, contrary to the requirement that $x \neq 0$. By Theorem 4.2.1 on Page 65,

$\det (M - \lambda I) = 0$. \hfill (10.2)

(Equivalently you could write $\det (\lambda I - M) = 0$.) The expression, $\det (\lambda I - M)$ or equivalently, $\det (M - \lambda I)$ is a polynomial called the **characteristic polynomial** and the above equation is called the characteristic equation. For $M$ an $n \times n$ matrix, it follows from the theorem on expanding a matrix by its cofactor that $\det (M - \lambda I)$ is a polynomial of degree $n$. As such, the equation, (10.2) has a solution, $\lambda \in \mathbb{C}$ by the fundamental theorem of algebra. Is it actually an eigenvalue? The answer is yes and this follows from Observation 9.2.6 on Page 173 along with Theorem 4.2.1 on Page 65. Since $\det (M - \lambda I) = 0$ the matrix, $\det (M - \lambda I)$ cannot be one to one and so there exists a nonzero vector, $x$ such that $(M - \lambda I)x = 0$. This proves the following corollary.

**Corollary 10.1.3** Let $M$ be an $n \times n$ matrix and $\det (M - \lambda I) = 0$. Then there exists a nonzero vector, $x \in \mathbb{C}^n$ such that $(M - \lambda I)x = 0$.

### 10.1.2 Finding Eigenvectors And Eigenvalues

As an example, consider the following.

**Example 10.1.4** Find the eigenvalues and eigenvectors for the matrix,

$$A = \begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix}.$$  

You first need to identify the eigenvalues. Recall this requires the solution of the equation

$$\det (A - \lambda I) = 0.$$  

In this case this equation is

$$\det \left( \begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) = 0.$$  

When you expand this determinant and simplify, you find the equation you need to solve is

$$(\lambda - 5) (\lambda^2 - 20\lambda + 100) = 0$$

and so the eigenvalues are

$$5, 10, 10.$$  

We have listed 10 twice because it is a zero of multiplicity two due to

$$\lambda^2 - 20\lambda + 100 = (\lambda - 10)^2.$$  

Having found the eigenvalues, it only remains to find the eigenvectors. First find the eigenvectors for $\lambda = 5$. As explained above, this requires you to solve the equation,

$$\begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} - 5 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$
That is you need to find the solution to
\[
\begin{pmatrix}
0 & -10 & -5 \\
2 & 9 & 2 \\
-4 & -8 & 1
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
=
\begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}
\]

By now this is an old problem. You set up the augmented matrix and row reduce to get the solution. Thus the matrix you must row reduce is
\[
\begin{pmatrix}
0 & -10 & -5 & | & 0 \\
2 & 9 & 2 & | & 0 \\
-4 & -8 & 1 & | & 0
\end{pmatrix}
\]
\[\text{(10.3)}\]

A reduced echelon form is
\[
\begin{pmatrix}
1 & 0 & -\frac{5}{4} & | & 0 \\
0 & 1 & \frac{1}{2} & | & 0 \\
0 & 0 & 0 & | & 0
\end{pmatrix}
\]

and so the solution is any vector of the form
\[
\begin{pmatrix}
\frac{5}{4}t \\
\frac{1}{2}t \\
t
\end{pmatrix}
= t
\begin{pmatrix}
\frac{5}{4} \\
\frac{1}{2} \\
1
\end{pmatrix}
\]

where \( t \in \mathbb{F} \). You would obtain the same collection of vectors if you replaced \( t \) with \( 4t \). Thus a simpler description for the solutions to this system of equations whose augmented matrix is in (10.3) is
\[
t
\begin{pmatrix}
5 \\
-2 \\
4
\end{pmatrix}
\]
\[\text{(10.4)}\]

where \( t \in \mathbb{F} \). Now you need to remember that you can’t take \( t = 0 \) because this would result in the zero vector and

**Eigenvectors are never equal to zero!**

Other than this value, every other choice of \( z \) in (10.4) results in an eigenvector. It is a good idea to check your work! To do so, we will take the original matrix and multiply by this vector and see if we get 5 times this vector.
\[
\begin{pmatrix}
5 & -10 & -5 \\
2 & 14 & 2 \\
-4 & -8 & 6
\end{pmatrix}
\begin{pmatrix}
5 \\
-2 \\
4
\end{pmatrix}
= 5
\begin{pmatrix}
5 \\
-2 \\
4
\end{pmatrix}
\]

so it appears this is correct. Always check your work on these problems if you care about getting the answer right.

The parameter, \( t \) is sometimes called a **free variable**. The set of vectors in (10.4) is called the **eigenspace** and it equals \( \ker (A - \lambda I) \). You should observe that in this case the eigenspace has dimension 1 because the eigenspace is the span of a single vector. In general, you obtain the solution from the row echelon form and the number of different free variables gives you the dimension of the eigenspace. Just remember that not every vector in the eigenspace is an eigenvector. The vector, \( 0 \) is not an eigenvector although it is in the eigenspace because

**Eigenvectors are never equal to zero!**
Next consider the eigenvectors for $\lambda = 10$. These vectors are solutions to the equation,

$$
\begin{pmatrix}
5 & -10 & -5 \\
2 & 14 & 2 \\
-4 & -8 & 6
\end{pmatrix}
- 10
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= 
\begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}
$$

That is you must find the solutions to

$$
\begin{pmatrix}
-5 & -10 & -5 \\
2 & 4 & 2 \\
-4 & -8 & -4
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= 
\begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}
$$

which reduces to consideration of the augmented matrix,

$$
\begin{pmatrix}
-5 & -10 & -5 & | & 0 \\
2 & 4 & 2 & | & 0 \\
-4 & -8 & 4 & | & 0
\end{pmatrix}
$$

A reduced echelon form for this matrix is

$$
\begin{pmatrix}
1 & 2 & 1 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
$$

and so the eigenvectors are of the form

$$
\begin{pmatrix}
-2s - t \\
s \\
t
\end{pmatrix}
= 
\begin{pmatrix}
-2 \\
1 \\
0
\end{pmatrix}
+ 
\begin{pmatrix}
-1 \\
0 \\
1
\end{pmatrix}.
$$

You can’t pick $t$ and $s$ both equal to zero because this would result in the zero vector and

**Eigenvectors are never equal to zero!**

However, every other choice of $t$ and $s$ does result in an eigenvector for the eigenvalue $\lambda = 10$. As in the case for $\lambda = 5$ you should check your work if you care about getting it right.

$$
\begin{pmatrix}
5 & -10 & -5 \\
2 & 14 & 2 \\
-4 & -8 & 6
\end{pmatrix}
\begin{pmatrix}
-1 \\
0 \\
1
\end{pmatrix}
= 
\begin{pmatrix}
-10 \\
0 \\
10
\end{pmatrix}
= 10
\begin{pmatrix}
-1 \\
0 \\
1
\end{pmatrix}
$$

so it worked. The other vector will also work. Check it.

**10.1.3 A Warning**

The above example shows how to find eigenvectors and eigenvalues algebraically. You may have noticed it is a bit long. Sometimes students try to first row reduce the matrix before looking for eigenvalues. This is a **terrible idea** because row operations destroy the eigenvalues. The eigenvalue problem is really not about row operations.

The general eigenvalue problem is the hardest problem in algebra and people still do research on ways to find eigenvalues and their eigenvectors. If you are doing anything which would yield a way to find eigenvalues and eigenvectors for general matrices without too much trouble, the thing you are doing will certainly be wrong. The problems you will see in these notes are not too hard because they are cooked up by us to be easy. Later we
will describe general methods to compute eigenvalues and eigenvectors numerically. These methods work even when the problem is not cooked up to be easy.

If you are so fortunate as to find the eigenvalues as in the above example, then finding the eigenvectors does reduce to row operations and this part of the problem is easy. However, finding the eigenvalues along with the eigenvectors is anything but easy because for an \( n \times n \) matrix, it involves solving a polynomial equation of degree \( n \). If you only find a good approximation to the eigenvalue, it won’t work. It either is or is not an eigenvalue and if it is not, the only solution to the equation, \((M - \lambda I)x = 0\) will be the zero solution as explained above and

\[
\text{Eigenvectors are never equal to zero!}
\]

Here is another example.

**Example 10.1.5** Let

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

First find the eigenvalues.

\[
\det \left( \begin{pmatrix}
2 & 2 & -2 \\
1 & 3 & -1 \\
-1 & 1 & 1
\end{pmatrix} - \lambda \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix} \right) = 0
\]

This reduces to \( \lambda^3 - 6\lambda^2 + 8\lambda = 0 \) and the solutions are 0, 2, and 4.

\[
\text{0 Can be an Eigenvalue!}
\]

Now find the eigenvectors. For \( \lambda = 0 \) the augmented matrix for finding the solutions is

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

and the a reduced echelon form is

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

Therefore, the eigenvectors are of the form

\[
t \begin{pmatrix}
1 \\
0 \\
1
\end{pmatrix}
\]

where \( t \neq 0 \).

Next find the eigenvectors for \( \lambda = 2 \). The augmented matrix for the system of equations needed to find these eigenvectors is

\[
\begin{pmatrix}
0 & 2 & -2 & | & 0 \\
1 & 1 & -1 & | & 0 \\
-1 & 1 & -1 & | & 0
\end{pmatrix}
\]
and the a reduced echelon form is
\[
\begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & -1 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]
and so the eigenvectors are of the form
\[
t \begin{pmatrix}
0 \\
1 \\
1
\end{pmatrix}
\]
where \( t \neq 0 \).

Finally find the eigenvectors for \( \lambda = 4 \). The augmented matrix for the system of equations needed to find these eigenvectors is
\[
\begin{pmatrix}
-2 & 2 & -2 & | & 0 \\
1 & -1 & -1 & | & 0 \\
-1 & 1 & -3 & | & 0
\end{pmatrix}
\]
and a reduced echelon form is
\[
\begin{pmatrix}
1 & -1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\].
Therefore, the eigenvectors are of the form
\[
t \begin{pmatrix}
1 \\
1 \\
0
\end{pmatrix}
\]
where \( t \neq 0 \).

### 10.1.4 Complex Eigenvalues

Sometimes you have to consider eigenvalues which are complex numbers. This occurs in differential equations for example. You do these problems exactly the same way as you do the ones in which the eigenvalues are real. Here is an example.

**Example 10.1.6** Find the eigenvalues and eigenvectors of the matrix
\[
A = \begin{pmatrix}
1 & 0 & 0 \\
0 & 2 & -1 \\
0 & 1 & 2
\end{pmatrix}.
\]

You need to find the eigenvalues. Solve
\[
\det \left( \begin{pmatrix}
1 & 0 & 0 \\
0 & 2 & -1 \\
0 & 1 & 2
\end{pmatrix} - \lambda \begin{pmatrix}
1 & 0 \\
0 & 1 \\
0 & 1
\end{pmatrix} \right) = 0.
\]
This reduces to \((\lambda - 1)(\lambda^2 - 4\lambda + 5) = 0\). The solutions are \( \lambda = 1, \lambda = 2 + i, \lambda = 2 - i \).

There is nothing new about finding the eigenvectors for \( \lambda = 1 \) so consider the eigenvalue \( \lambda = 2 + i \). You need to solve
\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
- \begin{pmatrix}
1 & 0 & 0 \\
0 & 2 & -1 \\
0 & 1 & 2
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
\]
In other words, you must consider the augmented matrix,

\[
\begin{pmatrix}
1 + i & 0 & 0 & | & 0 \\
0 & i & 1 & | & 0 \\
0 & -1 & i & | & 0
\end{pmatrix}
\]

for the solution. Divide the top row by \((1 + i)\) and then take \(-i\) times the second row and add to the bottom. This yields

\[
\begin{pmatrix}
1 & 0 & 0 & | & 0 \\
0 & i & 1 & | & 0 \\
0 & 0 & 0 & | & 0
\end{pmatrix}
\]

Now multiply the second row by \(-i\) to obtain

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

Therefore, the eigenvectors are of the form

\[
t \begin{pmatrix}
0 \\
i \\
1
\end{pmatrix}.
\]

You should find the eigenvectors for \(\lambda = 2 - i\). These are

\[
t \begin{pmatrix}
0 \\
-i \\
1
\end{pmatrix}.
\]

As usual, if you want to get it right you had better check it.

\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 2 & -1 \\
0 & 1 & 2
\end{pmatrix}
\begin{pmatrix}
0 \\
-i \\
1
\end{pmatrix} = \begin{pmatrix}
0 \\
-1 & -2i \\
2 & -i
\end{pmatrix} = (2 - i) \begin{pmatrix}
0 \\
-i \\
1
\end{pmatrix}
\]

so it worked.

### 10.2 Volumes

The determinant and the concept of eigenvalues and eigenvectors provide a way to give a unified treatment of the concept of volumes in various dimensions. First here is a useful theorem which is of considerable interest for its own sake.

**Theorem 10.2.1** Let \(A\) be an \(n \times n\) matrix. Then

\[
\det (A) = \prod_{i=1}^{n} \lambda_i
\]

where \(\lambda_i\) are the eigenvalues of \(A\). In words, the determinant of a matrix equals the product of its eigenvalues.
10.2. VOLUMES

Proof: The characteristic polynomial is \( \det (\lambda I - A) = \prod_{j=1}^{n} (\lambda - \lambda_j) \) where \( \lambda_j \) are the eigenvalues. This follows from the fundamental theorem of algebra which says every polynomial can be factored. Then, letting \( \lambda = 0 \) it follows \( \det (-A) = (-1)^n \det (A) = \prod_{j=1}^{n} (0 - \lambda_j) = (-1)^n \prod_{j=1}^{n} \lambda_j \) and this proves the theorem.

Recall the geometric definition of the cross product of two vectors found on Page 143. As explained there, the magnitude of the cross product of two vectors was the area of the parallelogram determined by the two vectors. There was also a coordinate description of the cross product. In terms of the notation of Proposition 7.5.4 on Page 153 the \( i \)th coordinate of the cross product is given by

\[
\varepsilon_{ijk} u_j v_k
\]

where the two vectors are \((u_1, u_2, u_3)\) and \((v_1, v_2, v_3)\). Therefore, using the reduction identity of Lemma 7.5.3 on Page 153 the \( i \)th coordinate

\[
|u \times v|^2 = \varepsilon_{ijk} u_j v_k \varepsilon_{irs} u_r v_s
\]

which equals

\[
\det \begin{pmatrix} u \cdot u & u \cdot v & u \cdot w \\ v \cdot u & v \cdot v & v \cdot w \\ w \cdot u & w \cdot v & w \cdot w \end{pmatrix}.
\]

Now recall the box product and how the box product was \( \pm \) the volume of the parallelepiped spanned by the three vectors. From the definition of the box product

\[
u \times v \cdot w = \begin{vmatrix} i & j & k \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} \cdot (w_1 i + w_2 j + w_3 k)
\]

\[= \det \begin{pmatrix} w_1 & w_2 & w_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{pmatrix}.
\]

Therefore,

\[
|u \times v \cdot w|^2 = \det \begin{pmatrix} w_1 & w_2 & w_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{pmatrix}^2
\]

which from the theory of determinants equals

\[
\det \begin{pmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix} \det \begin{pmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ u_3 & v_3 & w_3 \end{pmatrix} =
\]

\[
\det \begin{pmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix} \det \begin{pmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ u_3 & v_3 & w_3 \end{pmatrix} =
\]

\[
\det \begin{pmatrix} u_1^2 + v_1^2 + w_1^2 & u_1v_1 + u_2v_2 + u_3v_3 & u_1w_1 + u_2w_2 + u_3w_3 \\ u_1v_1 + u_2v_2 + u_3v_3 & v_1^2 + v_2^2 + v_3^2 & v_1w_1 + v_2w_2 + v_3w_3 \\ u_1w_1 + u_2w_2 + u_3w_3 & v_1w_1 + v_2w_2 + v_3w_3 & w_1^2 + w_2^2 + w_3^2 \end{pmatrix}
\]

\[= \det \begin{pmatrix} u \cdot u & u \cdot v & u \cdot w \\ u \cdot v & v \cdot v & v \cdot w \\ u \cdot w & v \cdot w & w \cdot w \end{pmatrix}.
\]
You see there is a definite pattern emerging here. These earlier cases were for a parallelepiped determined by either two or three vectors in $\mathbb{R}^3$. It makes sense to speak of a parallelepiped in any number of dimensions.

**Definition 10.2.2** Let $\mathbf{u}_1, \ldots, \mathbf{u}_p$ be vectors in $\mathbb{R}^k$. 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

$$\text{volume of } P(\mathbf{u}_1, \ldots, \mathbf{u}_p) \equiv (\det (\mathbf{u}_i \cdot \mathbf{u}_j))^{1/2}.$$ 

In this definition, $\mathbf{u}_i \cdot \mathbf{u}_j$ is the $ij$th entry of a $p \times p$ matrix. Note this definition agrees with all earlier notions of area and volume for parallelepipeds and it makes sense in any number of dimensions. However, it is important to verify the above determinant is nonnegative. After all, the above definition requires a square root of this determinant.

**Lemma 10.2.3** Let $\mathbf{u}_1, \ldots, \mathbf{u}_p$ be vectors in $\mathbb{R}^k$ for some $k$. Then $\det (\mathbf{u}_i \cdot \mathbf{u}_j) \geq 0$.

**Proof:** Recall $\mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T \mathbf{w}$. Therefore, in terms of matrix multiplication, the matrix $(\mathbf{u}_i \cdot \mathbf{u}_j)$ is just the following

$$ \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_p^T \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{pmatrix} $$

which is of the form $U^T U$.

Now the eigenvalues of the matrix $U^T U$ are all nonnegative. Here is why. Suppose $U^T U \mathbf{x} = \lambda \mathbf{x}$. Then

$$0 \leq \mathbf{x}^T U^T U \mathbf{x} = \lambda \mathbf{x}^T \mathbf{x} = \lambda \sum_k |x_k|^2.$$ 

Therefore, from Theorem 10.2.1, $\det (U^T U) \geq 0$ because it is the product of nonnegative numbers. This proves the lemma and shows the definition of volume is well defined.

Note it gives the right answer in the case where all the vectors are perpendicular. Here is why. Suppose $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ are vectors which have the property that $\mathbf{u}_i \cdot \mathbf{u}_j = 0$ if $i \neq j$. Thus $P(\mathbf{u}_1, \ldots, \mathbf{u}_p)$ is a box which has all $p$ sides perpendicular. What should its volume be? Shouldn’t it equal the product of the lengths of the sides? What does $\det (\mathbf{u}_i \cdot \mathbf{u}_j)$ give? The matrix $(\mathbf{u}_i \cdot \mathbf{u}_j)$ is a diagonal matrix having the squares of the magnitudes of the sides down the diagonal. Therefore, $\det (\mathbf{u}_i \cdot \mathbf{u}_j)^{1/2}$ equals the product of the lengths of the sides as it should. The matrix, $(\mathbf{u}_i \cdot \mathbf{u}_j)$ whose determinant gives the square of the volume of the parallelepiped spanned by the vectors, $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ is called the Grammian matrix and sometimes the metric tensor.

These considerations are of great significance because they allow the computation in a systematic manner of $k$ dimensional volumes of parallelepipeds which happen to be in $\mathbb{R}^n$ for $n \neq k$. Think for example of a plane in $\mathbb{R}^3$ and the problem of finding the area of something on this plane.
Example 10.2.4 Find the equation of the plane containing the three points, \((1,2,3)\), \((0,2,1)\), and \((3,1,0)\).

These three points determine two vectors, the one from \((0,2,1)\) to \((1,2,3)\), \(1 + 0j + 2k\), and the one from \((0,2,1)\) to \((3,1,0)\), \(3i + (-1)j + (-1)k\). If \((x,y,z)\) denotes a point in the plane, then the volume of the parallelepiped spanned by the vector from \((0,2,1)\) to \((x,y,z)\) and these other two vectors must be zero. Thus

\[
\begin{vmatrix}
x & y - 2 & z - 1 \\
3 & -1 & -1 \\
1 & 0 & 2 \\
\end{vmatrix} = 0
\]

Therefore, \(-2x - 7y + 13 + z = 0\) is the equation of the plane. You should check it contains all three points.

10.3 Block Multiplication Of Matrices

Suppose \(A\) is a matrix of the form

\[
\begin{pmatrix}
A_{11} & \cdots & A_{1m} \\
\vdots & \ddots & \vdots \\
A_{r1} & \cdots & A_{rm}
\end{pmatrix}
\]

where each \(A_{ij}\) is itself a matrix. Such a matrix is called a block matrix.

Suppose also that \(B\) is also a block matrix of the form

\[
\begin{pmatrix}
B_{11} & \cdots & B_{1p} \\
\vdots & \ddots & \vdots \\
B_{m1} & \cdots & B_{mp}
\end{pmatrix}
\]

and that for all \(i, j\), it makes sense to multiply \(A_{ik}B_{kj}\) for all \(k \in \{1, \ldots, m\}\) (That is the two matrices are conformable,) and that for each \(k\), \(A_{ik}B_{kj}\) is the same size. Then you can obtain \(AB\) as a block matrix as follows. \(AB = C\) where \(C\) is a block matrix having \(r \times p\) blocks such that the \(ij^{th}\) block is of the form

\[
C_{ij} = \sum_{k=1}^{m} A_{ik}B_{kj}.
\]

This is just like matrix multiplication for matrices whose entries are scalars except you must worry about the order of the factors and you sum matrices rather than numbers. Why should this formula hold? If \(m = 1\), the product is of the form

\[
\begin{pmatrix}
A_{11} \\
\vdots \\
A_{r1}
\end{pmatrix}
\times
\begin{pmatrix}
B_{11} & \cdots & B_{1p} \\
B_{m1} & \cdots & B_{mp}
\end{pmatrix}
\]

Considering the columns of \(B\) and the rows of \(A\) in the usual manner, (Recall the \(ij^{th}\) scalar entry of \(AB\) equals the dot product of the \(i^{th}\) row of \(A\) with the \(j^{th}\) column of \(B\).) it follows the formula will hold in this case. If \(m = 2\), the problem is of the form

\[
\begin{pmatrix}
A_{11} & A_{12} \\
\vdots & \vdots \\
A_{r1} & A_{r2}
\end{pmatrix}
\times
\begin{pmatrix}
B_{11} & \cdots & B_{1p} \\
B_{21} & \cdots & B_{2p}
\end{pmatrix}
\]

However, letting $\widetilde{A}_j = ( A_{j1} \ A_{j2} )$ and $\widetilde{B}_{ij} = \begin{pmatrix} B_{1j} \\ B_{2j} \end{pmatrix}$ then this appears in the form

$$\begin{pmatrix} \widetilde{A}_{11} \\ \vdots \\ \widetilde{A}_{r1} \end{pmatrix} \begin{pmatrix} \widetilde{B}_{11} & \cdots & \widetilde{B}_{1p} \end{pmatrix}$$

Now it is also not hard to see that

$$\widetilde{A}_j \widetilde{B}_{is} = \begin{pmatrix} A_{j1} \ A_{j2} \end{pmatrix} \begin{pmatrix} B_{1s} \\ B_{2s} \end{pmatrix} = (A_{j1}B_{1s} + A_{j2}B_{2s}) = \sum_{i=1}^{2} A_{ji}B_{is}$$

and by the first part, this would be the $js^{th}$ entry. Continuing this way, verifies the assertion about block multiplication. The reader is invited to give a more complete proof if desired.

This simple idea of block multiplication turns out to be very useful later. For now here is an interesting and significant application. In this theorem, $p_M(t)$ denotes the polynomial, $\det (tI - M)$. Thus the zeros of this polynomial are the eigenvalues of the matrix, $M$.

**Theorem 10.3.1** Let $A$ be an $m \times n$ matrix and let $B$ be an $n \times m$ matrix for $m \leq n$. Then

$$p_{BA}(t) = t^{n-m}p_{AB}(t),$$

so the eigenvalues of $BA$ and $AB$ are the same including multiplicities except that $BA$ has $n - m$ extra zero eigenvalues.

**Proof:** Use block multiplication to write

$$\begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} AB & ABA \\ B & BA \end{pmatrix}$$

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} AB & 0 \\ B & BA \end{pmatrix} = \begin{pmatrix} AB & ABA \\ B & BA \end{pmatrix}.$$ 

Therefore,

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} AB & 0 \\ B & BA \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}.$$ 

By Problem 12 of Page 89, it follows that $\begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}$ and $\begin{pmatrix} AB & 0 \\ B & BA \end{pmatrix}$ have the same characteristic polynomials. Therefore, noting that $BA$ is an $n \times n$ matrix and $AB$ is an $m \times m$ matrix,

$$t^m \det(tI - BA) = t^n \det(tI - AB)$$

and so $\det(tI - BA) = p_{BA}(t) = t^{n-m} \det(tI - AB) = t^{n-m}p_{AB}(t)$. This proves the theorem.

### 10.4 Shur’s Theorem

Every matrix is related to an upper triangular matrix in a particularly significant way. This is Shur’s theorem and it is the most important theorem in the spectral theory of matrices. Recall the Gram Schmidt procedure of Lemma 8.1.3 on Page 158 which is stated here for convenience.
**Lemma 10.4.1** Let \( \{x_1, \cdots, x_n\} \) be a basis for \( \mathbb{F}^n \). Then there exists an orthonormal basis for \( \mathbb{F}^n \), \( \{u_1, \cdots, u_n\} \) which has the property that for each \( k \leq n \), \( \text{span}(x_1, \cdots, x_k) = \text{span}(u_1, \cdots, u_k) \).

**Definition 10.4.2** 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 \( U_{ij} = U_{ji}^* \).

**Theorem 10.4.3** Let \( A \) be an \( n \times n \) matrix. Then there exists a unitary matrix, \( U \) such that

\[
U^* AU = T, \quad (10.5)
\]

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.

**Proof:** 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 procedure to obtain \( \{v_1, \cdots, v_n\} \), an orthonormal basis in \( \mathbb{F}^n \). Let \( U_0 \) be a matrix whose \( i^{th} \) column is \( v_i \). Then from the above, it follows \( U_0 \) is unitary. Then \( U_0^* AU_0 \) is of the form

\[
\begin{pmatrix}
\lambda_1 & * & \cdots & * \\
0 & & & \\
\vdots & & A_1 & \\
0 & & & \\
\end{pmatrix}
\]

where \( A_1 \) is an \( n-1 \times n-1 \) matrix. Repeat the process for the matrix, \( A_1 \) above. There exists a unitary matrix \( \tilde{U}_1 \) such that \( \tilde{U}_1^* A_1 \tilde{U}_1 \) is of the form

\[
\begin{pmatrix}
\lambda_2 & * & \cdots & * \\
0 & & & \\
\vdots & & A_2 & \\
0 & & & \\
\end{pmatrix}
\]

Now let \( U_1 \) be the \( n \times n \) matrix of the form

\[
\begin{pmatrix}
1 & 0 \\
0 & \tilde{U}_1
\end{pmatrix}.
\]

This is also a unitary matrix because by block multiplication,

\[
\begin{pmatrix}
1 & 0 \\
0 & \tilde{U}_1
\end{pmatrix}^* \begin{pmatrix}
1 & 0 \\
0 & \tilde{U}_1
\end{pmatrix} = \begin{pmatrix}
1 & 0 & 1 & 0 \\
0 & \tilde{U}_1^* & 0 & \tilde{U}_1
\end{pmatrix} = \begin{pmatrix}
1 & 0 \\
0 & \tilde{U}_1^* \tilde{U}_1
\end{pmatrix} = \begin{pmatrix}
1 & 0 \\
0 & I
\end{pmatrix}
\]

Then using block multiplication, \( U_1^* U_0^* AU_0 U_1 \) is of the form

\[
\begin{pmatrix}
\lambda_1 & * & \cdots & * \\
0 & \lambda_2 & * & \cdots & * \\
0 & 0 & & & \\
\vdots & \vdots & & A_2 & \\
0 & 0 & & & \\
\end{pmatrix}
\]
where $A_2$ is an $n - 2 \times n - 2$ matrix. Continuing in this way, there exists a unitary matrix, $U$ given as the product of the $U_i$ in the above construction such that

$$U^*AU = T$$

where $T$ is some upper triangular matrix. Since the matrix is upper triangular, the characteristic equation is $\prod_{i=1}^n (\lambda - \lambda_i)$ where the $\lambda_i$ are the diagonal entries of $T$. Therefore, the $\lambda_i$ are the eigenvalues.

What if $A$ is a real matrix and you only want to consider real unitary matrices?

**Theorem 10.4.4** Let $A$ be a real $n \times n$ matrix. Then there exists a real unitary matrix, $Q$ and a matrix $T$ of the form

$$T = \begin{pmatrix} P_1 & \cdots & \ast \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_r \end{pmatrix}$$

(10.6)

where $P_i$ equals either a real $1 \times 1$ matrix or $P_i$ equals a real $2 \times 2$ matrix having two complex eigenvalues of $A$ such that $Q^T AQ = T$. The matrix, $T$ is called a real Schur form of the matrix $A$.

**Proof:** Suppose

$$AV_1 = \lambda_1 V_1, \ |V_1| = 1$$

where $\lambda_1$ is real. Then let $\{v_1, \ldots, v_n\}$ be an orthonormal basis of vectors in $\mathbb{R}^n$. Let $Q_0$ be a matrix whose $i^{th}$ column is $v_i$. Then $Q_0^T AQ_0$ is of the form

$$\begin{pmatrix} \lambda_1 & \ast & \cdots & \ast \\ 0 & \ddots & \vdots & \vdots \\ \vdots & \ddots & A_1 & \ast \\ 0 & \cdots & 0 & A_2 \end{pmatrix}$$

where $A_1$ is a real $n - 1 \times n - 1$ matrix. This is just like the proof of Theorem 10.4.3 up to this point.

Now in case $\lambda_1 = \alpha + i\beta$, it follows since $A$ is real that $v_1 = z_1 + iw_1$ and that $\overline{v}_1 = z_1 - iw_1$ is an eigenvector for the eigenvalue, $\alpha - i\beta$. Here $z_1$ and $w_1$ are real vectors. It is clear that \{z_1, w_1\} is an independent set of vectors in $\mathbb{R}^n$. Indeed, \{v_1, \overline{v}_1\} is an independent set and it follows span (v_1, \overline{v}_1) = span (z_1, w_1). Now using the Gram Schmidt theorem in $\mathbb{R}^n$, there exists \{u_1, u_2\}, an orthonormal set of real vectors such that span (u_1, u_2) = span (v_1, \overline{v}_1).

Now let $\{u_1, u_2, \ldots, u_n\}$ be an orthonormal basis in $\mathbb{R}^n$ and let $Q_0$ be a unitary matrix whose $i^{th}$ column is $u_i$. Then $A u_j$ are both in span (u_1, u_2) for $j = 1, 2$ and so $u_k^T A u_j = 0$ whenever $k \geq 3$. It follows that $Q_0^T AQ_0$ is of the form

$$\begin{pmatrix} \ast & \ast & \cdots & \ast \\ \ast & \ast & \cdots & \ast \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & A_1 & \ast \\ 0 & \cdots & 0 & A_2 \end{pmatrix}$$

where $A_1$ is now an $n - 2 \times n - 2$ matrix. In this case, find $\tilde{Q}_1$ an $n - 2 \times n - 2$ matrix to put $A_1$ in an appropriate form as above and come up with $A_2$ either an $n - 4 \times n - 4$ matrix
or an \( n - 3 \times n - 3 \) matrix. Then the only other difference is to let

\[
Q_1 = \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & \ddots & \ddots & \ddots \\
\vdots & \vdots & \ddots & \ddots & \ddots \\
0 & 0 & \cdots & 1 & \tilde{Q}_1
\end{pmatrix}
\]

thus putting a \( 2 \times 2 \) identity matrix in the upper left corner rather than a one. Repeating this process with the above modification for the case of a complex eigenvalue leads eventually to (10.6) where \( Q \) is the product of real unitary matrices \( Q_i \) above. Finally,

\[
\lambda I - T = \begin{pmatrix}
\lambda & P_1 & \cdots & \ast \\
& \ddots & \ddots & \ddots \\
& & \ddots & \ddots \\
0 & \cdots & \lambda & P_r
\end{pmatrix}
\]

where \( I_k \) is the \( 2 \times 2 \) identity matrix in the case that \( P_k \) is \( 2 \times 2 \) and is the number 1 in the case where \( P_k \) is a \( 1 \times 1 \) matrix. Now, it follows that \( \det (\lambda I - T) = \prod_{k=1}^{r} \det (\lambda I_k - P_k) \). Therefore, \( \lambda \) is an eigenvalue of \( T \) if and only if it is an eigenvalue of some \( P_k \). This proves the theorem since the eigenvalues of \( T \) are the same as those of \( A \) because they have the same characteristic polynomial due to the similarity of \( A \) and \( T \).

**Definition 10.4.5** When a linear transformation, \( A \), mapping a linear space, \( V \) to \( V \) has a basis of eigenvectors, the linear transformation is called non defective. Otherwise it is called defective. An \( n \times n \) matrix, \( A \), is called normal if \( AA^* = A^*A \). An important class of normal matrices is that of the Hermitian or self adjoint matrices. An \( n \times n \) matrix, \( A \) is self adjoint or Hermitian if \( A = A^* \).

The next lemma is the basis for concluding that every normal matrix is unitarily similar to a diagonal matrix.

**Lemma 10.4.6** If \( T \) is upper triangular and normal, then \( T \) is a diagonal matrix.

**Proof:** Since \( T \) is normal, \( T^*T = TT^* \). Writing this in terms of components and using the description of the adjoint as the transpose of the conjugate, yields the following for the \( ik^{th} \) entry of \( T^*T = TT^* \).

\[
\sum_j t_{ij}^*t_{jk} = \sum_j \overline{t_{ij}}t_{kj} = \sum_j t_{ij}^*t_{jk} = \sum_j \overline{t_{ji}}t_{jk}.
\]

Now use the fact that \( T \) is upper triangular and let \( i = k = 1 \) to obtain the following from the above.

\[
\sum_j |t_{1j}|^2 \leq \sum_j |t_{1j}|^2 = |t_{11}|^2 \]

You see, \( t_{1j} = 0 \) unless \( j = 1 \) due to the assumption that \( T \) is upper triangular. This shows \( T \) is of the form

\[
\begin{pmatrix}
* & 0 & \cdots & 0 \\
0 & * & \cdots & * \\
\vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & *
\end{pmatrix}
\]
Now do the same thing only this time take \( i = k = 2 \) and use the result just established. Thus, from the above,

\[
\sum_j |t_{2j}|^2 = \sum_j |t_{j2}|^2 = |t_{22}|^2,
\]

showing that \( t_{2j} = 0 \) if \( j > 2 \) which means \( T \) has the form

\[
\begin{pmatrix}
  * & 0 & 0 & \cdots & 0 \\
  0 & * & 0 & \cdots & 0 \\
  0 & 0 & * & \cdots & * \\
  \vdots & \vdots & \ddots & \ddots & \vdots \\
  0 & 0 & 0 & 0 & * 
\end{pmatrix}.
\]

Next let \( i = k = 3 \) and obtain that \( T \) looks like a diagonal matrix in so far as the first 3 rows and columns are concerned. Continuing in this way it follows \( T \) is a diagonal matrix.

**Theorem 10.4.7** Let \( A \) be a normal matrix. Then there exists a unitary matrix, \( U \) such that \( U^*AU \) is a diagonal matrix.

**Proof:** From Theorem 10.4.3 there exists a unitary matrix, \( U \) such that \( U^*AU \) equals an upper triangular matrix. The theorem is now proved if it is shown that the property of being normal is preserved under unitary similarity transformations. That is, verify that if \( A \) is normal and if \( B = U^*AU \), then \( B \) is also normal. But this is easy.

\[
B^*B = U^*A^*UU^*AU = U^*A^*AU = U^*AA^*U = U^*AUU^*A^*U = BB^*.
\]

Therefore, \( U^*AU \) is a normal and upper triangular matrix and by Lemma 10.4.6 it must be a diagonal matrix. This proves the theorem.

**Corollary 10.4.8** If \( A \) is Hermitian, then all the eigenvalues of \( A \) are real and there exists an orthonormal basis of eigenvectors.

**Proof:** Since \( A \) is normal, there exists unitary, \( U \) such that \( U^*AU \) equals a diagonal matrix whose diagonal entries are the eigenvalues of \( A \). Therefore, \( D^* = U^*A^*U = U^*AU = D \) showing \( D \) is real.

Finally, let

\[
U = \begin{pmatrix} u_1 & u_2 & \cdots & u_n \end{pmatrix}
\]

where the \( u_i \) denote the columns of \( U \) and

\[
D = \begin{pmatrix} \lambda_1 & 0 \\
0 & \ddots \\
0 & \ddots & \lambda_n \end{pmatrix}
\]

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 \( ij \)th entry of \( U^*U \) equals \( \delta_{ij} \) and so

\[
\delta_{ij} = u_i^* u_j = \overline{u_i^* u_j} = u_i \cdot u_j.
\]

This proves the corollary because it shows the vectors \( \{u_i\} \) form an orthonormal basis.
Corollary 10.4.9 If \( A \) is a real symmetric matrix, then \( A \) is Hermitian and there exists a real unitary matrix, \( U \) such that \( U^T A U = D \) where \( D \) is a diagonal matrix.

Proof: This follows from Theorem 10.4.4 and Corollary 10.4.8. Alternatively, you could use Corollary 10.4.8 to assert the eigenvalues are all real. Then if \( A x = \lambda x \) the same is true of \( x \) and so in the construction for Shur’s theorem, you can always deal exclusively with real eigenvectors as long as your matrices are real and symmetric. When you construct the matrix which reduces the problem to a smaller one having \( A_1 \) in the lower right corner, use the Gram Schmidt process on \( \mathbb{R}^n \) using the real dot product to construct vectors, \( v_2, \ldots, v_n \) in \( \mathbb{R}^n \) such that \( \{v_1, \ldots, v_n\} \) is an orthonormal basis for \( \mathbb{R}^n \). The matrix \( A_1 \) is symmetric also. This is because for \( j, k \geq 2 \)

\[
A_{1kj} = v_k^T A v_j = (v_k^T A v_j)^T = v_j^T A v_k = A_{1jk}.
\]

Therefore, continuing this way, the process of the proof delivers only real vectors and real matrices.

10.5 Exercises

1. State the eigenvalue problem from an algebraic perspective.
2. State the eigenvalue problem from a geometric perspective.
3. Suppose \( T \) is a linear transformation and it satisfies \( T^2 = T \) and \( T x = x \) for all \( x \) in a certain subspace, \( V \). Show that 1 is an eigenvalue for \( T \) and show that all eigenvalues have absolute values no larger than 1.
4. Is it possible for a nonzero matrix to have only 0 as an eigenvalue?
5. Show that if \( A x = \lambda x \) and \( A y = \lambda y \), then whenever \( a, b \) are scalars,

\[
A (ax + by) = \lambda (ax + by).
\]

Does this imply that \( ax + by \) is an eigenvector? Explain.
6. Let \( M \) be an \( n \times n \) matrix and suppose \( x_1, \ldots, x_n \) are \( n \) eigenvectors which form a linearly independent set. Form the matrix \( S \) by making the columns these vectors. Show that \( S^{-1} MS \) exists and that \( S^{-1} MS \) is a diagonal matrix (one having zeros everywhere except on the main diagonal) having the eigenvalues of \( M \) on the main diagonal. When this can be done the matrix is diagonalizable.
7. Show that a matrix, \( M \) is diagonalizable if and only if it has a basis of eigenvectors.

Hint: The first part is done in Problem 6. It only remains to show that if the matrix can be diagonalized by some matrix, \( S \) giving \( D = S^{-1} MS \) for \( D \) a diagonal matrix, then it has a basis of eigenvectors. Try using the columns of the matrix \( S \).
8. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
-19 & -14 & -1 \\
8 & 4 & 8 \\
15 & 30 & -3
\end{pmatrix}.
\]
9. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
-3 & -30 & 15 \\
0 & 12 & 0 \\
15 & 30 & -3
\end{pmatrix}
\].

10. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
8 & 4 & 5 \\
0 & 12 & 9 \\
-2 & 2 & 10
\end{pmatrix}
\].

11. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
7 & -2 & 0 \\
8 & -1 & 0 \\
-2 & 4 & 6
\end{pmatrix}
\].

Can you find three independent eigenvectors?

12. Find the eigenvalues and eigenvectors of the matrix

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

Can you find three independent eigenvectors in this case?

13. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
12 & -12 & 6 \\
0 & 18 & 0 \\
6 & 12 & 12
\end{pmatrix}
\].

14. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
-5 & -1 & 10 \\
-15 & 9 & -6 \\
8 & -8 & 2
\end{pmatrix}
\].

15. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
-10 & -8 & 8 \\
-4 & -14 & -4 \\
0 & 0 & -18
\end{pmatrix}
\].

16. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
1 & 26 & -17 \\
4 & -4 & 4 \\
-9 & -18 & 9
\end{pmatrix}
\].

17. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
8 & 4 & 5 \\
0 & 12 & 9 \\
-2 & 2 & 10
\end{pmatrix}
\].
10.5. EXERCISES

18. Find the eigenvalues and eigenvectors of the matrix
\[
\begin{pmatrix}
9 & 6 & -3 \\
0 & 6 & 0 \\
-3 & -6 & 9
\end{pmatrix}.
\]

19. Find the eigenvalues and eigenvectors of the matrix
\[
\begin{pmatrix}
-10 & -2 & 11 \\
-18 & 6 & -9 \\
10 & -10 & -2
\end{pmatrix}.
\]

20. Find the complex eigenvalues and eigenvectors of the matrix
\[
\begin{pmatrix}
4 & -2 & -2 \\
0 & 2 & -2 \\
2 & 0 & 2
\end{pmatrix}.
\]

21. Let \( T \) be the linear transformation which reflects vectors about the \( x \) axis. Find a matrix for \( T \) and then find its eigenvalues and eigenvectors.

22. Let \( T \) be the linear transformation which rotates all vectors in \( \mathbb{R}^2 \) counterclockwise through an angle of \( \pi/2 \). Find a matrix of \( T \) and then find eigenvalues and eigenvectors.

23. Let \( T \) be the linear transformation which reflects all vectors in \( \mathbb{R}^3 \) through the \( xy \) plane. Find a matrix for \( T \) and then obtain its eigenvalues and eigenvectors.

24. 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 volume of the parallelepiped determined by these three vectors.

25. 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.

26. Here are three vectors in \( \mathbb{R}^2 : (1, 2)^T, (2, 1)^T, (0, 1)^T \). Find the volume of the parallelepiped determined by these three vectors. Why should this volume equal zero?

27. If there are \( n + 1 \) or more vectors in \( \mathbb{R}^n \), Lemma 10.2.3 implies the parallelepiped determined by these \( n + 1 \) vectors must have zero volume. What is the geometric significance of this assertion?

28. Find the equation of the plane through the three points \((1, 2, 3), (2, -3, 1), (1, 1, 7)\).

29. Let
\[
A = \begin{pmatrix}
1 & 2 \\
3 & 4 \\
0 & 1
\end{pmatrix}
\quad \text{and let}
\]
\[
B = \begin{pmatrix}
0 & 1 \\
1 & 1 \\
2 & 1
\end{pmatrix}
\]
Multiply \( AB \) verifying the block multiplication formula. Here \( A_{11} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \), \( A_{12} = \begin{pmatrix} 2 \\ 0 \end{pmatrix} \), \( A_{21} = \begin{pmatrix} 0 & 1 \end{pmatrix} \) and \( A_{22} = (3) \).
Let $A$ be an $r \times r$ matrix and let $B$ be an $m \times m$ matrix such that $r + m = n$. Consider the following $n \times n$ block matrix
\[
C = \begin{pmatrix}
A & 0 \\
D & B
\end{pmatrix},
\]
where the $D$ is an $m \times r$ matrix, and the $0$ is a $r \times m$ matrix. Letting $I_k$ denote the $k \times k$ identity matrix, tell why
\[
C = \begin{pmatrix}
A & 0 \\
D & I_m
\end{pmatrix}\begin{pmatrix}
I_r & 0 \\
0 & B
\end{pmatrix}.
\]
Now explain why $\det(C) = \det(A) \det(B)$. **Hint:** Part of this will require an explanation of why
\[
\det \begin{pmatrix}
A & 0 \\
D & I_m
\end{pmatrix} = \det(A).
\]
See Theorems 4.1.23 - 4.1.25.

31. If $A$ is a real $n \times n$ matrix which has all real eigenvalues, show there exists a real unitary matrix, $U$ such that $U^TAU = T$ where $T$ is a real upper triangular matrix. If $A$ is normal, explain why $T$ is a diagonal matrix.

32. If $A$ is an $n \times n$ Hermitian matrix, show there exists an orthonormal basis, \( \{v_1, \cdots, v_n\} \) such that
\[
A = \sum_{j=1}^{n} \lambda_j v_j^T v_j.
\]
If $A$ is real and Hermitian, show that the vectors, $v_j$ may all be taken to be real vectors.
Planes, And Surfaces In $\mathbb{R}^n$

11.0.1 Outcomes

1. Find the angle between two lines.

2. Determine a point of intersection between a line and a surface.

3. Find the equation of a plane in 3 space given a point and a normal vector, three points, a sketch of a plane or a geometric description of the plane.

4. Determine the normal vector and the intercepts of a given plane.

5. Sketch the graph of a plane given its equation.

6. Determine the cosine of the angle between two planes.

7. Find the equation of a plane determined by lines.

8. Identify standard quadric surfaces given their functions or graphs.

9. Sketch the graph of a quadric surface by identifying the intercepts, traces, sections, symmetry and boundedness or unboundedness of the surface.

11.1 Planes

You have an idea of what a plane is already. It is the span of some vectors. However, it can also be considered geometrically in terms of a dot product. To find the equation of a plane, you need two things, a point contained in the plane and a vector normal to the plane. Let $\mathbf{p}_0 = (x_0, y_0, z_0)$ denote the position vector of a point in the plane, let $\mathbf{p} = (x, y, z)$ be the position vector of an arbitrary point in the plane, and let $\mathbf{n}$ denote a vector normal to the plane. This means that

$$\mathbf{n} \cdot (\mathbf{p} - \mathbf{p}_0) = 0$$

whenever $\mathbf{p}$ is the position vector of a point in the plane. The following picture illustrates the geometry of this idea.
Expressed equivalently, the plane is just the set of all points \( p \) such that the vector, \( p - p_0 \) is perpendicular to the given normal vector, \( n \).

**Example 11.1.1** Find the equation of the plane with normal vector, \( n = (1, 2, 3) \) containing the point \((2, -1, 5)\).

From the above, the equation of this plane is just
\[
(1, 2, 3) \cdot (x - 2, y + 1, z - 3) = x - 9 + 2y + 3z = 0
\]

**Example 11.1.2** \(2x + 4y - 5z = 11\) is the equation of a plane. Find the normal vector and a point on this plane.

You can write this in the form \(2 \begin{pmatrix} x - \frac{11}{2} \end{pmatrix} + 4 \begin{pmatrix} y - 0 \end{pmatrix} + \begin{pmatrix} -5 \end{pmatrix} \begin{pmatrix} z - 0 \end{pmatrix} = 0\). Therefore, a normal vector to the plane is \(2i + 4j - 5k\) and a point in this plane is \(\left(\frac{11}{2}, 0, 0\right)\). Of course there are many other points in the plane.

**Definition 11.1.3** Suppose two planes intersect. The angle between the planes is defined to be the angle between their normal vectors.

**Example 11.1.4** Find the equation of the plane which contains the three points, \((1, 2, 1)\), \((3, -1, 2)\), and \((4, 2, 1)\) another way.

You just need to get a normal vector to this plane. This can be done by taking the cross products of the two vectors,
\[
(3, -1, 2) - (1, 2, 1) \quad \text{and} \quad (4, 2, 1) - (1, 2, 1)
\]
Thus a normal vector is \((2, -3, 1) \times (3, 0, 0) = (0, 3, 9)\). Therefore, the equation of the plane is
\[
0(x - 1) + 3(y - 2) + 9(z - 1) = 0
\]
or \(3y + 9z = 15\) which is the same as \(y + 3z = 5\).

**Example 11.1.5** Find the equation of the plane which contains the three points, \((1, 2, 1)\), \((3, -1, 2)\), and \((4, 2, 1)\) another way.

Letting \((x, y, z)\) be a point on the plane, the volume of the parallelepiped spanned by \((x, y, z) - (1, 2, 1)\) and the two vectors, \((2, -3, 1)\) and \((3, 0, 0)\) must be equal to zero. Thus the equation of the plane is
\[
\det \begin{pmatrix} 3 & 0 & 0 \\ 2 & -3 & 1 \\ x - 1 & y - 2 & z - 1 \end{pmatrix} = 0.
\]
Hence \(-9z + 15 - 3y = 0\) and dividing by 3 yields the same answer as the above.
Proposition 11.1.6 If \((a, b, c) \neq (0, 0, 0)\), then \(ax + by + cz = d\) is the equation of a plane with normal vector \(ai + bj + ck\). Conversely, any plane can be written in this form.

Proof: One of \(a, b, c\) is nonzero. Suppose for example that \(c \neq 0\). Then the equation can be written as

\[
a(x - 0) + b(y - 0) + c\left(z - \frac{d}{c}\right) = 0
\]

Therefore, \((0, 0, \frac{d}{c})\) is a point on the plane and a normal vector is \(ai + bj + ck\). The converse follows from the above discussion involving the point and a normal vector. This proves the proposition.

Example 11.1.7 Find the equation of the plane which contains the three points, \((1, 2, 1)\), \((3, -1, 2)\), and \((4, 2, 1)\) another way.

You need to find numbers, \(a, b, c, d\) not all zero such that each of the given three points satisfies the equation,

\[
ax + by + cz = d.
\]

Then you must have for \((x, y, z)\) a point on this plane,

\[
\begin{align*}
a + 2b + c - d &= 0, \\
3a - b + 2c - d &= 0, \\
4a + 2b + c - d &= 0, \\
xa + yb + zc - d &= 0.
\end{align*}
\]

You need a nonzero solution to the above system of four equations for the unknowns, \(a, b, c,\) and \(d\). Therefore,

\[
\det\begin{pmatrix}
1 & 2 & 1 & -1 \\
3 & -1 & 2 & -1 \\
4 & 2 & 1 & -1 \\
x & y & z & -1
\end{pmatrix} = 0
\]

because the matrix sends a nonzero vector, \((a, b, c, -d)\) to zero and is therefore, not one to one. Consequently from Theorem 4.2.1 on Page 65, its determinant equals zero. Hence upon evaluating the determinant,

\[-15 + 9z + 3y = 0\]

which reduces to \(3z + y = 5\).

Example 11.1.8 Find the equation of the plane containing the points \((1, 2, 3)\) and the line \((0, 1, 1) + t(2, 1, 2) = (x, y, z)\).

There are several ways to do this. One is to find three points and use any of the above procedures. Let \(t = 0\) and then let \(t = 1\) to get two points on the line. This yields \((1, 2, 3), (0, 1, 1),\) and \((2, 2, 3)\). Then the equation of the plane is

\[
\det\begin{pmatrix}
x & y & z & -1 \\
1 & 2 & 3 & -1 \\
0 & 1 & 1 & -1 \\
2 & 2 & 3 & -1
\end{pmatrix} = 2y - z - 1 = 0.
\]

Example 11.1.9 Find the equation of the plane which contains the two lines, given by the following parametric expressions in which \(t \in \mathbb{R}\).

\[
(2t, 1 + t, 1 + 2t) = (x, y, z), \quad (2t + 2, 1, 3 + 2t) = (x, y, z)
\]
Note first that you don’t know there even is such a plane. However, if there is, you could find it by obtaining three points, two on one line and one on another and then using any of the above procedures for finding the plane. From the first line, two points are (0,1,1) and (2,2,3) while a third point can be obtained from second line, (2,1,3). You need a normal vector and then use any of these points. To get a normal vector, form $(2,0,2) \times (2,1,2) = (-2,0,2)$. Therefore, the plane is $-2x+0(y-1)+2(z-1)=0$. This reduces to $z-x=1$.

If there is a plane, this is it. Now you can simply verify that both of the lines are really in this plane. From the first, $(1+2t)-2t=1$ and the second, $(3+2t)-(2t+2)=1$ so both lines lie in the plane.

One way to understand how a plane looks is to connect the points where it intercepts the $x$, $y$, and $z$ axes. This allows you to visualize the plane somewhat and is a good way to sketch the plane. Not surprisingly these points are called intercepts.

**Example 11.1.10** Sketch the plane which has intercepts $(2,0,0)$, $(0,3,0)$, and $(0,0,4)$.

You see how connecting the intercepts gives a fairly good geometric description of the plane. These lines which connect the intercepts are also called the traces of the plane. Thus the line which joins $(0,3,0)$ to $(0,0,4)$ is the intersection of the plane with the $yz$ plane. It is the trace on the $yz$ plane.

**Example 11.1.11** Identify the intercepts of the plane, $3x - 4y + 5z = 11$.

The easy way to do this is to divide both sides by 11.

$$
\frac{x}{(11/3)} + \frac{y}{(-11/4)} + \frac{z}{(11/5)} = 1
$$

The intercepts are $(11/3,0,0)$, $(0, -11/4,0)$ and $(0,0,11/5)$. You can see this by letting both $y$ and $z$ equal to zero to find the point on the $x$ axis which is intersected by the plane. The other axes are handled similarly.

### 11.2 Quadric Surfaces

In the above it was shown that the equation of an arbitrary plane is an equation of the form $ax + by + cz = d$. Such equations are called level surfaces. There are some standard level surfaces which involve certain variables being raised to a power of 2 which are sufficiently important that they are given names, usually involving the portentous semi-word “oid”. These are graphed below using Maple, a computer algebra system.
Why do the graphs of these level surfaces look the way they do? Consider first the
hyperboloid of two sheets. The equation defining this surface can be written in the form
\[ \frac{z^2}{a^2} - 1 = \frac{x^2}{b^2} + \frac{y^2}{c^2}. \]

Suppose you fix a value for \( z \). What ordered pairs, \((x, y)\) will satisfy the equation? If \( \frac{z^2}{a^2} < 1 \), there is no such ordered pair because the above equation would require a negative number to equal a nonnegative one. This is why there is a gap and there are two sheets. If \( \frac{z^2}{a^2} > 1 \), then the above equation is the equation for an ellipse. That is why if you slice the graph by letting \( z = z_0 \) the result is an ellipse in the plane \( z = z_0 \).

Consider the hyperboloid of one sheet.
\[ \frac{x^2}{b^2} + \frac{y^2}{c^2} = 1 + \frac{z^2}{a^2}. \]

This time, it doesn’t matter what value \( z \) takes. The resulting equation for \((x, y)\) is an ellipse.

Similar considerations apply to the elliptic paraboloid as long as \( z > 0 \) and the ellipsoid.

The elliptic cone is like the hyperboloid of two sheets without the 1. Therefore, \( z \) can have any value. In case \( z = 0 \), \((x, y) = (0, 0)\). Viewed from the side, it appears straight, not curved like the hyperboloid of two sheets. This is because if \((x, y, z)\) is a point on the surface, then if \( t \) is a scalar, it follows \((tx, ty, tz)\) is also on this surface.

The most interesting of these graphs is the hyperbolic paraboloid
\[ 1 + \frac{z^2}{a^2}, \] which opens to the right and left while if \( z < 0 \) it is a hyperbola which opens up and down. As \( z \) passes from positive to negative, the hyperbola changes type and this is what yields the shape shown in the picture.

Not surprisingly, you can find intercepts and traces of quadric surfaces just as with planes.

**Example 11.2.1** Find the trace on the \( xy \) plane of the hyperbolic paraboloid, \( z = x^2 - y^2 \).

This occurs when \( z = 0 \) and so this reduces to \( y^2 = x^2 \). In other words, this trace is just the two straight lines, \( y = x \) and \( y = -x \).

**Example 11.2.2** Find the intercepts of the ellipsoid, \( x^2 + 2y^2 + 4z^2 = 9 \).

To find the intercept on the \( x \) axis, let \( y = z = 0 \) and this yields \( x = \pm 3 \). Thus there are two intercepts, \((3, 0, 0)\) and \((-3, 0, 0)\). The other intercepts are left for you to find. You can see this is an aid in graphing the quadric surface. The surface is said to be bounded if there is some number, \( C \) such that whenever, \((x, y, z)\) is a point on the surface, \( \sqrt{x^2 + y^2 + z^2} < C \). The surface is called unbounded if no such constant, \( C \) exists. Ellipsoids are bounded but the other quadric surfaces are not bounded.

**Example 11.2.3** Why is the hyperboloid of one sheet, \( x^2 + 2y^2 - z^2 = 1 \) unbounded?

Let \( z \) be very large. Does there correspond \((x, y)\) such that \((x, y, z)\) is a point on the hyperboloid of one sheet? Certainly. Simply pick any \((x, y)\) on the ellipse \( x^2 + 2y^2 = 1 + z^2 \). Then \( \sqrt{x^2 + y^2 + z^2} \) is large, at lest as large as \( z \). Thus it is unbounded.

You can also find intersections between lines and surfaces.

**Example 11.2.4** Find the points of intersection of the line \((x, y, z) = (1 + t, 1 + 2t, 1 + t)\) with the surface, \( z = x^2 + y^2 \).

\(^1\)It is traditional to refer to this as a hyperbolic paraboloid. Not a parabolic hyperboloid.
First of all, there is no guarantee there is any intersection at all. But if it exists, you have only to solve the equation for \( t \)

\[
1 + t = (1 + t)^2 + (1 + 2t)^2
\]

This occurs at the two values of \( t = -\frac{1}{2} + \frac{1}{10}\sqrt{5}, t = -\frac{1}{2} - \frac{1}{10}\sqrt{5}. \) Therefore, the two points are

\[
(1, 1, 1) + \left( -\frac{1}{2} + \frac{1}{10}\sqrt{5} \right) (1, 2, 1), \quad \text{and} \quad (1, 1, 1) + \left( -\frac{1}{2} - \frac{1}{10}\sqrt{5} \right) (1, 2, 1)
\]

That is

\[
\left( \frac{1}{2} + \frac{1}{10}\sqrt{5}, \frac{1}{5}\sqrt{5}, \frac{1}{2} + \frac{1}{10}\sqrt{5} \right), \left( \frac{1}{2} - \frac{1}{10}\sqrt{5}, -\frac{1}{5}\sqrt{5}, \frac{1}{2} - \frac{1}{10}\sqrt{5} \right).
\]

### 11.3 Exercises

1. Determine whether the lines \((1, 1, 2) + t(1, 0, 3)\) and \((4, 1, 3) + t(3, 0, 1)\) have a point of intersection. If they do, find the cosine of the angle between the two lines. If they do not intersect, explain why they do not.

2. Determine whether the lines \((1, 1, 2) + t(1, 0, 3)\) and \((4, 2, 3) + t(3, 0, 1)\) have a point of intersection. If they do, find the cosine of the angle between the two lines. If they do not intersect, explain why they do not.

3. Find where the line \((1, 0, 1) + t(1, 2, 1)\) intersects the surface \(x^2 + y^2 + z^2 = 9\) if possible.
   If there is no intersection, explain why.

4. Find a parametric equation for the line through the points \((2, 3, 4, 5)\) and \((-2, 3, 0, 1)\).

5. Find the equation of a line through \((1, 2, 3, 0)\) which has direction vector, \((2, 1, 3, 1)\).

6. Let \((x, y) = (2\cos(t), 2\sin(t))\) where \(t \in [0, 2\pi]\). Describe the set of points encountered as \(t\) changes.

7. Let \((x, y, z) = (2\cos(t), 2\sin(t), t)\) where \(t \in \mathbb{R}\). Describe the set of points encountered as \(t\) changes.

8. If there is a plane which contains the two lines, \((2t + 2, 1 + t, 3 + 2t) = (x, y, z)\) and \((4 + t, 3 + 2t, 4 + t) = (x, y, z)\) find it. If there is no such plane tell why.

9. If there is a plane which contains the two lines, \((2t + 4, 1 + t, 3 + 2t) = (x, y, z)\) and \((4 + t, 3 + 2t, 4 + t) = (x, y, z)\) find it. If there is no such plane tell why.

10. Find the equation of the plane which contains the three points \((1, -2, 3), (2, 3, 4), \) and \((3, 1, 2)\).

11. Find the equation of the plane which contains the three points \((1, 2, 3), (2, 0, 4), \) and \((3, 1, 2)\).

12. Find the equation of the plane which contains the three points \((0, 2, 3), (2, 3, 4), \) and \((3, 5, 2)\).

13. Find the equation of the plane which contains the three points \((1, 2, 3), (0, 3, 4), \) and \((3, 6, 2)\).

14. Find the equation of the plane having a normal vector, \(5\mathbf{i} + 2\mathbf{j} - 6\mathbf{k}\) which contains the point \((2, 1, 3)\).
15. Find the equation of the plane having a normal vector, $\mathbf{i} + 2\mathbf{j} - 4\mathbf{k}$ which contains the point $(2, 0, 1)$.

16. Find the equation of the plane having a normal vector, $2\mathbf{i} + \mathbf{j} - 6\mathbf{k}$ which contains the point $(1, 1, 2)$.

17. Find the equation of the plane having a normal vector, $\mathbf{i} + 2\mathbf{j} - 3\mathbf{k}$ which contains the point $(1, 0, 3)$.

18. Find the cosine of the angle between the two planes $2x + 3y - z = 11$ and $3x + y + 2z = 9$.

20. Find the cosine of the angle between the two planes $2x + y - z = 11$ and $3x + 5y + 2z = 9$.

21. Find the cosine of the angle between the two planes $x + 3y + z = 11$ and $3x + 2y + 2z = 9$.

22. Determine the intercepts and sketch the plane $3x - 2y + z = 4$.

23. Determine the intercepts and sketch the plane $x - 2y + z = 2$.

24. Determine the intercepts and sketch the plane $x + y + z = 3$.

25. Based on an analogy with the above pictures, sketch or otherwise describe the graph of $y = \frac{x^2}{a^2} - \frac{z^2}{b^2}$.

26. Based on an analogy with the above pictures, sketch or otherwise describe the graph of $\frac{x^2}{a^2} + \frac{y^2}{c^2} = 1 + \frac{z^2}{d^2}$.

27. Find the intercepts of the quadric surface, $x^2 + 4y^2 - z^2 = 4$ and sketch the surface.

28. Find the intercepts of the quadric surface, $x^2 - (4y^2 + z^2) = 4$ and sketch the surface.

29. Find the intersection of the line $(x, y, z) = (1 + t, t, 3t)$ with the surface, $x^2/9 + y^2/4 + z^2/16 = 1$ if possible.
Vector Valued Functions

12.0.1 Outcomes

1. Identify the domain of a vector function.

2. Represent combinations of multivariable functions algebraically.

3. Evaluate the limit of a function of several variables or show that it does not exist.

4. Determine whether a function is continuous at a given point. Give examples of continuous functions.

5. Recall and apply the extreme value theorem.

12.1 Vector Valued Functions

Vector valued functions have values in \( \mathbb{R}^p \) where \( p \) is an integer at least as large as 1. Here is a simple example which is obviously of interest.

Example 12.1.1 A rocket is launched from the rotating earth. You could define a function having values in \( \mathbb{R}^3 \) as \( (r(t), \theta(t), \phi(t)) \) where \( r(t) \) is the distance of the center of mass of the rocket from the center of the earth, \( \theta(t) \) is the longitude, and \( \phi(t) \) is the latitude of the rocket.

Example 12.1.2 Let \( f(x, y) = (\sin xy, y^3 + x, x^4) \). Then \( f \) is a function defined on \( \mathbb{R}^2 \) which has values in \( \mathbb{R}^3 \). For example, \( f(1, 2) = (\sin 2, 9, 16) \).

As usual, \( D(f) \) denotes the domain of the function, \( f \) which is written in bold face because it will possibly have values in \( \mathbb{R}^p \). When \( D(f) \) is not specified, it will be understood that the domain of \( f \) consists of those things for which \( f \) makes sense.

Example 12.1.3 Let \( f(x, y, z) = \left(\frac{x+y}{2}, \sqrt{1-x^2}, y \right) \). Then \( D(f) \) would consist of the set of all \( (x, y, z) \) such that \( |x| \leq 1 \) and \( z \neq 0 \).

There are many ways to make new functions from old ones.

Definition 12.1.4 Let \( f, g \) be functions with values in \( \mathbb{R}^p \). Let \( a, b \) be elements of \( \mathbb{R} \) (scalars). Then \( af + bg \) is the name of a function whose domain is \( D(f) \cap D(g) \) which is defined as

\[
(af + bg)(x) = af(x) + bg(x).
\]
f · g or \((f, g)\) is the name of a function whose domain is \(D(f) \cap D(g)\) which is defined as
\[
(f, g)(x) \equiv f(x) \cdot g(x).
\]

If \(f\) and \(g\) have values in \(\mathbb{R}^3\), define a new function, \(f \times g\) by
\[
f \times g(t) \equiv f(t) \times g(t).
\]

If \(f : D(f) \to X\) and \(g : X \to Y\), then \(g \circ f\) is the name of a function whose domain is
\[
\{x \in D(f) : f(x) \in D(g)\}
\]
which is defined as
\[
g \circ f(x) \equiv g(f(x)).
\]
This is called the composition of the two functions.

You should note that \(f(x)\) is not a function. It is the value of the function at the point, \(x\). The name of the function is \(f\). Nevertheless, people often write \(f(x)\) to denote a function and it doesn’t cause too many problems in beginning courses. When this is done, the variable, \(x\) should be considered as a generic variable free to be anything in \(D(f)\). I will use this slightly sloppy abuse of notation whenever convenient.

**Example 12.1.5** Let \(f(t) \equiv (t, 1 + t, 2)\) and \(g(t) \equiv (t^2, t, t)\). Then \(f \cdot g\) is the name of the function satisfying
\[
f \cdot g(t) = f(t) \cdot g(t) = t^3 + t + t^2 + 2t = t^3 + t^2 + 3t.
\]

Note that in this case is was assumed the domains of the functions consisted of all of \(\mathbb{R}\) because this was the set on which the two both made sense. Also note that \(f\) and \(g\) map \(\mathbb{R}\) into \(\mathbb{R}^3\) but \(f \cdot g\) maps \(\mathbb{R}\) into \(\mathbb{R}\).

**Example 12.1.6** Suppose \(f(t) = (2t, 1 + t^2)\) and \(g : \mathbb{R}^2 \to \mathbb{R}\) is given by \(g(x, y) \equiv x + y\). Then \(g \circ f : \mathbb{R} \to \mathbb{R}\) and
\[
g \circ f(t) = g(f(t)) = g(2t, 1 + t^2) = 1 + 2t + t^2.
\]

### 12.2 Vector Fields

Some people find it useful to try and draw pictures to illustrate a vector valued function. This can be a very useful idea in the case where the function takes points in \(D \subseteq \mathbb{R}^2\) and delivers a vector in \(\mathbb{R}^2\). For many points, \((x, y) \in D\), you draw an arrow of the appropriate length and direction with its tail at \((x, y)\). The picture of all these arrows can give you an understanding of what is happening. For example if the vector valued function gives the velocity of a fluid at the point, \((x, y)\), the picture of these arrows can give an idea of the motion of the fluid. When they are long the fluid is moving fast, when they are short, the fluid is moving slowly the direction of these arrows is an indication of the direction of motion. The only sensible way to produce such a picture is with a computer. Otherwise, it becomes a worthless exercise in busy work. Furthermore, it is of limited usefulness in three dimensions because in three dimensions such pictures are too cluttered to convey much insight.

**Example 12.2.1** Draw a picture of the vector field, \((-x, y)\) which gives the velocity of a fluid flowing in two dimensions.
Example 12.2.2 Draw a picture of the vector field \((y, x)\) for the velocity of a fluid flowing in two dimensions.

Example 12.2.3 Draw a picture of the vector field \((y \cos(x) + 1, x \sin(y) - 1)\) for the velocity of a fluid flowing in two dimensions.
12.3 Continuous Functions

What was done in beginning calculus for scalar functions is generalized here to include the case of a vector valued function.

**Definition 12.3.1** A function \( f : D ( f ) \subseteq \mathbb{R}^p \to \mathbb{R}^q \) is continuous at \( x \in D ( f ) \) if for each \( \varepsilon > 0 \) there exists \( \delta > 0 \) such that whenever \( y \in D ( f ) \) and \( |y - x| < \delta \) it follows that \( |f(x) - f(y)| < \varepsilon \).

\( f \) is continuous if it is continuous at every point of \( D ( f ) \).

Note the total similarity to the scalar valued case.

**12.3.1 Sufficient Conditions For Continuity**

The next theorem is a fundamental result which will allow us to worry less about the \( \varepsilon \delta \) definition of continuity.

**Theorem 12.3.2** The following assertions are valid.

1. The function, \( af + bg \) is continuous at \( x \) whenever \( f, g \) are continuous at \( x \in D ( f ) \cap D ( g ) \) and \( a, b \in \mathbb{R} \).

2. If \( f \) is continuous at \( x \), \( f(x) \in D ( g ) \subseteq \mathbb{R}^p \), and \( g \) is continuous at \( f(x) \), then \( g \circ f \) is continuous at \( x \).

3. If \( f = (f_1, \ldots, f_q) : D ( f ) \to \mathbb{R}^q \), then \( f \) is continuous if and only if each \( f_k \) is a continuous real valued function.

4. The function \( f : \mathbb{R}^p \to \mathbb{R} \), given by \( f(x) = |x| \) is continuous.

The proof of this theorem is in the last section of this chapter. Its conclusions are not surprising. For example the first claim says that \( (af + bg)(y) \) is close to \( (af + bg)(x) \) when \( y \) is close to \( x \) provided the same can be said about \( f \) and \( g \). For the second claim, if \( y \) is close to \( x \), \( f(x) \) is close to \( f(y) \) and so by continuity of \( g \) at \( f(x) \), \( g(f(y)) \) is close to \( g(f(x)) \). To see the third claim is likely, note that closeness in \( \mathbb{R}^p \) is the same as closeness in each coordinate. The fourth claim is immediate from the triangle inequality.

For functions defined on \( \mathbb{R}^n \), there is a notion of polynomial just as there is for functions defined on \( \mathbb{R} \).

**Definition 12.3.3** Let \( \alpha \) be an \( n \) dimensional multi-index. This means

\[ \alpha = (\alpha_1, \ldots, \alpha_n) \]

where each \( \alpha_i \) is a natural number or zero. Also, let

\[ |\alpha| = \sum_{i=1}^{n} |\alpha_i| \]

The symbol, \( x^\alpha \) means

\[ x^\alpha = x_1^{\alpha_1}x_2^{\alpha_2} \cdots x_n^{\alpha_n} \].
An $n$ dimensional polynomial of degree $m$ is a function of the form 

$$p(x) = \sum_{|\alpha| \leq m} d_\alpha x^\alpha.$$ 

where the $d_\alpha$ are real numbers.

The above theorem implies that polynomials are all continuous.

### 12.4 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.

**Definition 12.4.1** Let $A \subseteq \mathbb{R}^m$ 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 12.4.2** Let $f : D(f) \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^q$ be a function and let $x$ be a limit point of $D(f)$.

Then 

$$\lim_{y \rightarrow x} f(y) = L$$

if and only if the following condition holds. For all $\varepsilon > 0$ there exists $\delta > 0$ such that if $0 < |y - x| < \delta$ and $y \in D(f)$ then,

$$|L - f(y)| < \varepsilon.$$

**Theorem 12.4.3** If $\lim_{y \rightarrow x} f(y) = L$ and $\lim_{y \rightarrow x} f(y) = L_1$, 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

$$|f(y) - L| < \varepsilon, \quad |f(y) - L_1| < \varepsilon.$$

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 $L = L_1$.

As in the case of functions of one variable, one can define what it means for $\lim_{y \rightarrow x} f(x) = \pm \infty$.

**Definition 12.4.4** If $f(x) \in \mathbb{R}$, $\lim_{y \rightarrow x} f(x) = \infty$ if for every number $l$, there exists $\delta > 0$ such that whenever $|y - x| < \delta$ and $y \in D(f)$, then $f(x) > l$.

The following theorem is just like the one variable version of calculus.

**Theorem 12.4.5** Suppose $\lim_{y \rightarrow x} f(y) = L$ and $\lim_{y \rightarrow x} g(y) = K$ where $K, L \in \mathbb{R}^q$. Then if $a, b \in \mathbb{R}$,

$$\lim_{y \rightarrow x} (af(y) + bg(y)) = aL + bK,$$  \hspace{1cm} (12.1)

$$\lim_{y \rightarrow x} f \cdot g(y) = LK$$  \hspace{1cm} (12.2)
and if \( g \) is scalar valued with \( \lim_{y \to x} g(y) = K \neq 0 \),
\[
\lim_{y \to x} f(y) g(y) = LK. \tag{12.3}
\]

Also, if \( h \) is a continuous function defined near \( L \), then
\[
\lim_{y \to x} h \circ f(y) = h(L). \tag{12.4}
\]

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:** The proof of (12.1) is left for you. It is like a corresponding theorem for continuous functions. Now (12.2) is to be verified. Let \( \varepsilon > 0 \) be given. Then by the triangle inequality,
\[
|f \cdot g(y) - L \cdot K| \leq |f g(y) - f(y) \cdot K| + |f(y) \cdot K - L \cdot K| \\
\leq |f(y)||g(y) - K| + |K||f(y) - L|.
\]

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 \( 0 < |y - x| < \delta_1 \),
\[
|f \cdot g(y) - L \cdot K| \leq (1 + |K| + |L|)|g(y) - K| + |f(y) - L|. \tag{12.5}
\]

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 \( 0 < \delta = \min(\delta_1, \delta_2) \), it follows from (12.5) that
\[
|f \cdot g(y) - L \cdot K| < \varepsilon
\]
and this proves (12.2).

The proof of (12.3) is left to you.

Consider (12.4). 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

\[
|f(y) - L| = (\frac{\sum_{k=1}^{p} |g_k(y)|^2}{y} - L_k)^{1/2} \leq \sum_{k=1}^{p} |g_k(y) - L_k|^{1/2} = \sum_{k=1}^{p} \frac{\varepsilon^2}{p} = \varepsilon.
\]

The following theorem is important.

Theorem 12.4.7 Suppose \( f : D(f) \to \mathbb{R}^q \). Then for \( x \) a limit point of \( D(f) \),

\[
\lim_{y \to x} f(y) = L
\]

if and only if

\[
\lim_{y \to x} f_k(y) = L_k
\]

where \( f(y) = (f_1(y), \ldots, f_q(y)) \) and \( L = (L_1, \ldots, L_p) \).

In the case where \( q = 3 \) and \( \lim_{y \to x} f(y) = L \) and \( \lim_{y \to x} g(y) = K \), then

\[
\lim_{y \to x} f(y) \times g(y) = L \times K.
\]
It remains to verify (12.8). But from the first part of this theorem and the description of the cross product presented earlier in terms of the permutation symbol,

$$\lim_{y \to x} (f(y) \times g(y))_i = \lim_{y \to x} \epsilon_{ijk} f_j(y) g_k(y) = \epsilon_{ijk} L_i K_j = (L \times K)_i.$$ 

Therefore, from the first part of this theorem, this establishes (16.5). This completes the proof.

**Example 12.4.8** Find $$\lim_{(x,y) \to (3,1)} \left( x^2 - 9x - 3, y \right).$$

It is clear that $$\lim_{(x,y) \to (3,1)} x^2 - 9x - 3 = 6$$ and $$\lim_{(x,y) \to (3,1)} y = 1.$$ Therefore, this limit equals $$(6,1).$$

**Example 12.4.9** Find $$\lim_{(x,y) \to (0,0)} \frac{xy}{x^2 + y^2}.$$ 

First of all observe the domain of the function is $$\mathbb{R}^2 \setminus \{(0,0)\},$$ every point in $$\mathbb{R}^2$$ except the origin. Therefore, $$(0,0)$$ is a limit point of the domain of the function so it might make sense to take a limit. However, just as in the case of a function of one variable, the limit may not exist. In fact, this is the case here. To see this, take points on the line $$y = 0.$$ At these points, the value of the function equals 0. Now consider points on the line $$y = x$$ where the value of the function equals 1/2. Since arbitrarily close to $$(0,0)$$ there are points where the function equals 1/2 and points where the function has the value 0, it follows there can be no limit. Just take $$\varepsilon = 1/10$$ for example. You can’t be within 1/10 of 1/2 and also within 1/10 of 0 at the same time.

Note it is necessary to rely on the definition of the limit much more than in the case of a function of one variable and there are no easy ways to do limit problems for functions of more than one variable. It is what it is and you will not deal with these concepts without suffering and anguish.

### 12.5 Properties Of Continuous Functions

Functions of $$p$$ variables have many of the same properties as functions of one variable. First there is a version of the extreme value theorem generalizing the one dimensional case.

**Theorem 12.5.1** Let $$C$$ be closed and bounded and let $$f : C \to \mathbb{R}$$ be continuous. Then $$f$$ achieves its maximum and its minimum on $$C.$$ This means there exist, $$x_1, x_2 \in C$$ such that for all $$x \in C,$$

$$f(x_1) \leq f(x) \leq f(x_2).$$

There is also the long technical theorem about sums and products of continuous functions. These theorems are proved in the next section.

**Theorem 12.5.2** The following assertions are valid

1. The function, $$af + bg$$ is continuous at $$x$$ when $$f, g$$ are continuous at $$x \in D(f) \cap D(g)$$ and $$a, b \in \mathbb{R}.$$ 

2. If and $$f$$ and $$g$$ are each real valued functions continuous at $$x,$$ then $$fg$$ is continuous at $$x.$$ If, in addition to this, $$g(x) \neq 0,$$ then $$f/g$$ is continuous at $$x.$$
3. If $f$ is continuous at $x$, $f(x) \in D(g) \subseteq \mathbb{R}^p$, and $g$ is continuous at $f(x)$, then $g \circ f$ is continuous at $x$.

4. If $f = (f_1, \ldots, f_q) : D(f) \to \mathbb{R}^q$, then $f$ is continuous if and only if each $f_k$ is a continuous real valued function.

5. The function $f : \mathbb{R}^p \to \mathbb{R}$, given by $f(x) = |x|$ is continuous.

### 12.6 Exercises

1. Let $f(t) = (t, t^2 + 1, \frac{t}{t+1})$ and let $g(t) = (t + 1, 1, \frac{t}{t+1})$. Find $f \cdot g$.

2. Let $f, g$ be given in the previous problem. Find $f \times g$.

3. Find $D(f)$ if $f(x, y, z, w) = \left(\frac{xy}{zw}, \sqrt{6 - x^2y^2}\right)$.

4. Let $f(t) = (t, t^2, t^3)$, $g(t) = (1, t, t^2)$, and $h(t) = (\sin t, t, 1)$. Find the time rate of change of the volume of the parallelepiped spanned by the vectors $f, g$, and $h$.

5. Let $f(t) = (t, \sin t)$. Show $f$ is continuous at every point $t$.

6. Suppose $|f(x) - f(y)| \leq K|x - y|$ where $K$ is a constant. Show that $f$ is everywhere continuous. Functions satisfying such an inequality are called Lipschitz functions.

7. Suppose $|f(x) - f(y)| \leq K|x - y|^\alpha$ where $K$ is a constant and $\alpha \in (0, 1)$. Show that $f$ is everywhere continuous.

8. Suppose $f : \mathbb{R}^3 \to \mathbb{R}$ is given by $f(x) = 3x_1x_2 + 2x_3^2$. Use Theorem 12.3.2 to verify that $f$ is continuous. **Hint:** You should first verify that the function, $\pi_k : \mathbb{R}^3 \to \mathbb{R}$ given by $\pi_k(x) = x_k$ is a continuous function.

9. Show that if $f : \mathbb{R}^q \to \mathbb{R}$ is a polynomial then it is continuous.

10. State and prove a theorem which involves quotients of functions encountered in the previous problem.

11. Let

   \[ f(x, y) \equiv \begin{cases} \frac{xy}{x^2+y^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases} \]

   Find $\lim_{(x,y) \to (0,0)} f(x, y)$ if it exists. If it does not exist, tell why it does not exist. **Hint:** Consider along the line $y = x$ and along the line $y = 0$.

12. Find the following limits if possible

   (a) $\lim_{(x,y) \to (0,0)} \frac{x^2-y^2}{x^2+y^2}$

   (b) $\lim_{(x,y) \to (0,0)} \frac{xy}{x^2+y^2}$

   (c) $\lim_{(x,y) \to (0,0)} \frac{(x^2-y^2)^2}{x^2+y^2}$. **Hint:** Consider along $y = 0$ and along $x = y^2$.

   (d) $\lim_{(x,y) \to (0,0)} x \sin \left(\frac{1}{x+y^2}\right)$

   (e) $\lim_{(x,y) \to (1,2)} -2xy^2 + 8xy + 34y^3 - 18x^2 + 6x^2 - 13y - 20 - xy^2 - x^3$. **Hint:** It might help to write this in terms of the variables $(s, t) = (x - 1, y - 2)$.
13. In the definition of limit, why must \( x \) be a limit point of \( D (f) \)? **Hint:** If \( x \) were not a limit point of \( D (f) \), show there exists \( \delta > 0 \) such that \( B (x, \delta) \) contains no points of \( D (f) \) other than possibly \( x \) itself. Argue that 33.3 is a limit and that so is 22 and 7 and 11. In other words the concept is totally worthless.

14. Suppose \( \lim_{x \to 0} f(x,0) = 0 = \lim_{y \to 0} f(0,y) \). Does it follow that \( \lim_{(x,y) \to (0,0)} f(x,y) = 0 \)? Prove or give counter example.

15. **f :** \( D \subseteq \mathbb{R}^p \to \mathbb{R}^q \) is Lipschitz continuous or just Lipschitz for short if there exists a constant, \( K \) such that
\[
|f(x) - f(y)| \leq K|x - y|
\]
for all \( x, y \in D \). Show every Lipschitz function is uniformly continuous which means that given \( \varepsilon > 0 \) there exists \( \delta > 0 \) independent of \( x \) such that if \( |x - y| < \delta \), then \( |f(x) - f(y)| < \varepsilon \).

16. If \( f \) is uniformly continuous, does it follow that \( |f| \) is also uniformly continuous? If \( |f| \) is uniformly continuous does it follow that \( f \) is uniformly continuous? Answer the same questions with “uniformly continuous” replaced with “continuous”. Explain why.

17. Let \( f \) be defined on the positive integers. Thus \( D (f) = \mathbb{N} \). Show that \( f \) is automatically continuous at every point of \( D (f) \). Is it also uniformly continuous? What does this mean about the concept of continuous functions being those which can be graphed without taking the pencil off the paper?

18. In Problem 12c show \( \lim_{t \to 0} f(tx,ty) = 1 \) for any choice of \( (x,y) \). Using Problem 12c what does this tell you about limits existing just because the limit along any line exists.

19. Let \( f(x,y,z) = x^2y + \sin (xyz) \). Does \( f \) achieve a maximum on the set \( \{(x,y,z) : x^2 + y^2 + 2z^2 \leq 8\} \)? Explain why.

20. Suppose \( x \) is defined to be a limit point of a set, \( A \) if and only if for all \( r > 0 \), \( B (x, r) \) contains a point of \( A \) different than \( x \). Show this is equivalent to the above definition of limit point.

21. Give an example of a set of points in \( \mathbb{R}^3 \) which has no limit points. Show that if \( D (f) \) equals this set, then \( f \) is continuous. Show that more generally, if \( f \) is any function for which \( D (f) \) has no limit points, then \( f \) is continuous.

22. Let \( \{x_k\}_{k=1}^n \) be any finite set of points in \( \mathbb{R}^p \). Show this set has no limit points.

23. Suppose \( S \) is any set of points such that every pair of points is at least as far apart as 1. Show \( S \) has no limit points.

24. Find \( \lim_{x \to 0} \frac{\sin(|x|)}{|x|} \) and prove your answer from the definition of limit.

25. Suppose \( g \) is a continuous vector valued function of one variable defined on \([0, \infty)\). Prove
\[
\lim_{x \to x_0} g(|x|) = g(|x_0|).
\]

26. Give some examples of limit problems for functions of many variables which have limits and prove your assertions.
12.7 Some Fundamentals

This section contains the proofs of the theorems which were stated without proof along with some other significant topics which will be useful later. These topics are of fundamental significance but are difficult.

**Theorem 12.7.1** The following assertions are valid

1. The function, \( af + bg \) is continuous at \( x \) when \( f, g \) are continuous at \( x \in D(f) \cap D(g) \) and \( a, b \in \mathbb{R} \).
2. If and \( f \) and \( g \) are each real valued functions continuous at \( x \), then \( fg \) is continuous at \( x \). If, in addition to this, \( g(x) \neq 0 \), then \( f/g \) is continuous at \( x \).
3. If \( f \) is continuous at \( x \), \( f(x) \in D(g) \subseteq \mathbb{R}^p \), and \( g \) is continuous at \( f(x) \), then \( g \circ f \) is continuous at \( x \).
4. If \( f = (f_1, \ldots, f_q) : D(f) \to \mathbb{R}^q \), then \( f \) is continuous if and only if each \( f_k \) is a continuous real valued function.
5. The function \( f : \mathbb{R}^p \to \mathbb{R} \), given by \( f(x) = |x| \) is continuous.

**Proof:** Begin with 1.) Let \( \varepsilon > 0 \) be given. By assumption, there exist \( \delta_1 > 0 \) such that whenever \( |x - y| < \delta_1 \), it follows \( |f(x) - f(y)| < \varepsilon \frac{|a|}{2(|a| + |b| + 1)} \) and there exists \( \delta_2 > 0 \) such that whenever \( |x - y| < \delta_2 \), it follows \( |g(x) - g(y)| < \varepsilon \frac{1}{2(|a| + |b| + 1)} \). Then let \( 0 < \delta \leq \min(\delta_1, \delta_2) \). If \( |x - y| < \delta \), then everything happens at once. Therefore, using the triangle inequality

\[
|af(x) + bg(x) - (ag(y) + bg(y))| \\
\leq |a||f(x) - f(y)| + |b||g(x) - g(y)| \\
< |a| \left( \varepsilon \frac{1}{2(|a| + |b| + 1)} \right) + |b| \left( \varepsilon \frac{1}{2(|a| + |b| + 1)} \right) < \varepsilon.
\]

Now begin on 2.) There exists \( \delta_1 > 0 \) such that if \( |y - x| < \delta_1 \), then \( |f(x) - f(y)| < 1. \) Therefore, for such \( y \),

\[
|f(y)| < 1 + |f(x)|.
\]

It follows that for such \( y \),

\[
|fg(x) - fg(y)| \leq |f(x)g(x) - g(x)f(y)| + |g(x)f(y) - f(y)g(y)| \\
\leq |g(x)||f(x) - f(y)| + |f(y)||g(x) - g(y)| \\
\leq (1 + |g(x)| + |f(y)|)|g(x) - g(y)| + |f(x) - f(y)|.
\]
Now let $\varepsilon > 0$ be given. There exists $\delta_2$ such that if $|x - y| < \delta_2$, then

$$|g(x) - g(y)| < \frac{\varepsilon}{2(1 + |g(x)| + |f(y)|)},$$

and there exists $\delta_3$ such that if $|x - y| < \delta_3$, then

$$|f(x) - f(y)| < \frac{\varepsilon}{2(1 + |g(x)| + |f(y)|)}.$$

Now let $0 < \delta \leq \min(\delta_1, \delta_2, \delta_3)$. Then if $|x - y| < \delta$, all the above hold at once and

$$|f(g(x)) - f(g(y))| \leq (1 + |g(x)| + |f(y)|) |g(x) - g(y)| + |f(x) - f(y)|$$

$$\leq (1 + |g(x)| + |f(y)|) \left( \frac{\varepsilon}{2(1 + |g(x)| + |f(y)|)} + \frac{\varepsilon}{2(1 + |g(x)| + |f(y)|)} \right) = \varepsilon.$$

This proves the first part of 2.) To obtain the second part, let $\delta_1$ be as described above and let $\delta_0 > 0$ be such that for $|x - y| < \delta_0$,

$$|g(x) - g(y)| < |g(x)|/2$$

and so by the triangle inequality,

$$-|g(x)|/2 \leq |g(y)| - |g(x)| \leq |g(x)|/2$$

which implies $|g(y)| \geq |g(x)|/2$, and $|g(y)| < 3|g(x)|/2$.

Then if $|x - y| < \min(\delta_0, \delta_1)$,

$$\left| \frac{f(x)}{g(x)} - \frac{f(y)}{g(y)} \right| = \left| \frac{f(x)g(y) - f(y)g(x)}{g(x)g(y)} \right|$$

$$\leq \frac{2}{|g(x)|^2} \left| f(x)g(y) - f(y)g(x) \right|$$

$$\leq \frac{2}{|g(x)|^2} \left| g(y) \right| \left| f(x) - f(y) \right| + \frac{2}{|g(x)|^2} \left| f(y) \right| \left| g(y) - g(x) \right|$$

$$\leq \frac{2}{|g(x)|^2} \left[ \frac{3}{2} \left| g(x) \right| \left| f(x) - f(y) \right| + (1 + \left| f(x) \right|) \left| g(y) - g(x) \right| \right]$$

$$\leq \frac{2}{|g(x)|^2} (1 + 2|f(x)| + 2|g(x)|) \left| f(x) - f(y) \right| + \left| g(y) - g(x) \right|$$

$$\equiv M \left| f(x) - f(y) \right| + \left| g(y) - g(x) \right|$$

where

$$M = \frac{2}{|g(x)|^2} (1 + 2|f(x)| + 2|g(x)|)$$
Now let $\delta_2$ be such that if $|x - y| < \delta_2$, then
\[ |f(x) - f(y)| < \frac{\varepsilon}{2} M^{-1} \]
and let $\delta_3$ be such that if $|x - y| < \delta_3$, then
\[ |g(y) - g(x)| < \frac{\varepsilon}{2} M^{-1}. \]
Then if $0 < \delta \leq \min(\delta_0, \delta_1, \delta_2, \delta_3), \text{ and } |x - y| < \delta$, everything holds and
\[
\left| \frac{f(x)}{g(x)} - \frac{f(y)}{g(y)} \right| \leq M \left| f(x) - f(y) \right| + \left| g(y) - g(x) \right| \\
< M \left[ \frac{\varepsilon}{2} M^{-1} + \frac{\varepsilon}{2} M^{-1} \right] = \varepsilon.
\]
This completes the proof of the second part of 2.) Note that in these proofs no effort is made to find some sort of “best” $\delta$. The problem is one which has a yes or a no answer. Either is it or it is not continuous.

Now begin on 3.). If $f$ is continuous at $x$, $f(x) \in D(g) \subseteq \mathbb{R}^p$, and $g$ is continuous at $f(x)$, then $g \circ f$ is continuous at $x$. Let $\varepsilon > 0$ be given. Then there exists $\eta > 0$ such that if $|y - f(x)| < \eta$ and $y \in D(g)$, it follows that $|g(y) - g(f(x))| < \varepsilon$. It follows from continuity of $f$ at $x$ that there exists $\delta > 0$ such that if $|x - z| < \delta$ and $z \in D(f)$, then $|f(z) - f(x)| < \eta$. Then if $|x - z| < \delta$ and $z \in D(g \circ f) \subseteq D(f)$, all the above hold and so
\[ |g(f(z)) - g(f(x))| < \varepsilon. \]
This proves part 3.)

Part 4.) says: If $f = (f_1, \cdots, f_q) : D(f) \to \mathbb{R}^q$, then $f$ is continuous if and only if each $f_k$ is a continuous real valued function. Then
\[
|f_k(x) - f_k(y)| \leq |f(x) - f(y)| \\
= \left( \sum_{i=1}^q |f_i(x) - f_i(y)|^2 \right)^{1/2} \\
\leq \sum_{i=1}^q |f_i(x) - f_i(y)|. \quad (12.9)
\]
Suppose first that $f$ is continuous at $x$. Then there exists $\delta > 0$ such that if $|x - y| < \delta$, then $|f(x) - f(y)| < \varepsilon$. The first part of the above inequality then shows that for each $k = 1, \cdots, q$, $|f_k(x) - f_k(y)| < \varepsilon$. This shows the only if part. Now suppose each function, $f_k$ is continuous. Then if $\varepsilon > 0$ is given, there exists $\delta_k > 0$ such that whenever $|x - y| < \delta_k$
\[ |f_k(x) - f_k(y)| < \varepsilon/q. \]
Now let $0 < \delta \leq \min(\delta_1, \cdots, \delta_q)$. For $|x - y| < \delta$, the above inequality holds for all $k$ and so the last part of (12.9) implies
\[
|f(x) - f(y)| \leq \sum_{i=1}^q |f_i(x) - f_i(y)| \\
< \sum_{i=1}^q \frac{\varepsilon}{q} = \varepsilon.
\]
This proves part 4.)
To verify part 5.), let \( \varepsilon > 0 \) be given and let \( \delta = \varepsilon \). Then if \( |x - y| < \delta \), the triangle inequality implies
\[
|f(x) - f(y)| = |x| - |y| 
\leq |x - y| < \delta = \varepsilon.
\]
This proves part 5.) and completes the proof of the theorem.

12.7.1 The Nested Interval Lemma
Here is a multidimensional version of the nested interval lemma.

**Lemma 12.7.2** Let \( I_k = \prod_{i=1}^{p} [a^k_i, b^k_i] \equiv \{ x \in \mathbb{R}^p : x_i \in [a^k_i, b^k_i] \} \) and suppose that for all \( k = 1, 2, \cdots \),
\[
I_k \supseteq I_{k+1}.
\]
Then there exists a point, \( c \in \mathbb{R}^p \) which is an element of every \( I_k \).

**Proof:** Since \( I_k \supseteq I_{k+1} \), it follows that for each \( i = 1, \cdots, p \), \( [a^k_i, b^k_i] \supseteq [a^{k+1}_i, b^{k+1}_i] \).
This implies that for each \( i \),
\[
a^k_i \leq a^{k+1}_i, \quad b^k_i \geq b^{k+1}_i. \tag{12.10}
\]
Consequently, if \( k \leq l \),
\[
a^l_i \leq b^l_i \leq b^k_i. \tag{12.11}
\]
Now define
\[
c_i \equiv \sup \{ a^l_i : l = 1, 2, \cdots \}
\]
for each \( k = 1, 2, \cdots \). Therefore, picking any \( k \), (12.11) shows that \( b^k_i \) is an upper bound for the set \( \{ a^l_i : l = k, k+1, \cdots \} \) and so it is at least as large as the least upper bound of this set which is the definition of \( c_i \) given in (12.12). Thus, for each \( i \) and each \( k \),
\[
a^k_i \leq c_i \leq b^k_i.
\]
Defining \( c \equiv (c_1, \cdots, c_p) \), \( c \in I_k \) for all \( k \). This proves the lemma.

If you don’t like the proof, you could prove the lemma for the one variable case first and then do the following.

**Lemma 12.7.3** Let \( I_k = \prod_{i=1}^{p} [a^k_i, b^k_i] \equiv \{ x \in \mathbb{R}^p : x_i \in [a^k_i, b^k_i] \} \) and suppose that for all \( k = 1, 2, \cdots \),
\[
I_k \supseteq I_{k+1}.
\]
Then there exists a point, \( c \in \mathbb{R}^p \) which is an element of every \( I_k \).

**Proof:** For each \( i = 1, \cdots, p \), \( [a^k_i, b^k_i] \supseteq [a^{k+1}_i, b^{k+1}_i] \) and so by the nested interval theorem for one dimensional problems, there exists a point \( c_i \in [a^k_i, b^k_i] \) for all \( k \). Then letting \( c \equiv (c_1, \cdots, c_p) \) it follows \( c \in I_k \) for all \( k \). This proves the lemma.
12.7. SOME FUNDAMENTALS

12.7.2 The Extreme Value Theorem

Definition 12.7.4 A set, $C \subseteq \mathbb{R}^p$ is said to be bounded if $C \subseteq \prod_{i=1}^{p} [a_i, b_i]$ for some choice of intervals, $[a_i, b_i]$ where $-\infty < a_i < b_i < \infty$. The diameter of a set, $S$, is defined as

$$diam(S) \equiv \sup \{ |x - y| : x, y \in S \}.$$

A function, $f$ having values in $\mathbb{R}^p$ is said to be bounded if the set of values of $f$ is a bounded set.

Thus $diam(S)$ is just a careful description of what you would think of as the diameter. It measures how stretched out the set is.

Lemma 12.7.5 Let $C \subseteq \mathbb{R}^p$ be closed and bounded and let $f : C \to \mathbb{R}$ be continuous. Then $f$ is bounded.

Proof: Suppose not. Since $C$ is bounded, it follows $C \subseteq \prod_{i=1}^{p} [a_i, b_i] \equiv I_0$ for some closed intervals, $[a_i, b_i]$. Consider all sets of the form $\prod_{i=1}^{p} [c_i, d_i]$ where $[c_i, d_i]$ equals either $[a_i, a_i + b_i]$ or $[c_i, d_i] = [\frac{a_i + b_i}{2}, b_i]$. Thus there are $2^p$ of these sets because there are two choices for the $i^{th}$ slot for $i = 1, \cdots, p$. Also, if $x$ and $y$ are two points in one of these sets,

$$|x_i - y_i| \leq 2^{-1} |b_i - a_i|.$$

Observe that $diam(I_0) = \left(\sum_{i=1}^{p} |b_i - a_i|^2\right)^{1/2}$ because for $x, y \in I_0, |x_i - y_i| \leq |a_i - b_i|$ for each $i = 1, \cdots, p$,

$$|x - y| = \left(\sum_{i=1}^{p} |x_i - y_i|^2\right)^{1/2} \leq 2^{-1} \left(\sum_{i=1}^{p} |b_i - a_i|^2\right)^{1/2} \equiv 2^{-1} diam(I_0).$$

Denote by $\{J_1, \cdots, J_{2^p}\}$ these sets determined above. It follows the diameter of each set is no larger than $2^{-1} diam(I_0)$. In particular, since $d \equiv (d_1, \cdots, d_p)$ and $c \equiv (c_1, \cdots, c_p)$ are two such points, for each $J_k$,

$$diam(J_k) \equiv \left(\sum_{i=1}^{p} |d_i - c_i|^2\right)^{1/2} \leq 2^{-1} diam(I_0).$$

Since the union of these sets equals all of $I_0$, it follows

$$C = \cup_{k=1}^{2^p} J_k \cap C.$$

If $f$ is not bounded on $C$, it follows that for some $k$, $f$ is not bounded on $J_k \cap C$. Let $I_1 \equiv J_k$ and let $C_1 = C \cap I_1$. Now do to $I_1$ and $C_1$ what was done to $I_0$ and $C$ to obtain $I_2 \subseteq I_1$, and for $x, y \in I_2$,

$$|x - y| \leq 2^{-1} diam(I_1) \leq 2^{-2} diam(I_2),$$

and $f$ is unbounded on $I_2 \cap C_1 \equiv C_2$. Continue in this way obtaining sets, $I_k$ such that $I_k \supseteq I_{k+1}$ and $diam(I_k) \leq 2^{-k} diam(I_0)$ and $f$ is unbounded on $I_k \cap C$. By the nested interval lemma, there exists a point, $c$ which is contained in each $I_k$.

Claim: $c \in C$. 

Proof of claim: Suppose \( c \notin C \). Since \( C \) is a closed set, there exists \( r > 0 \) such that \( B(c, r) \) is contained completely in \( \mathbb{R}^p \setminus C \). In other words, \( B(c, r) \) contains no points of \( C \).

Let \( k \) be so large that \( \text{diam}(I_0)2^{-k} < r \). Then since \( c \in I_k \), and any two points of \( I_k \) are closer than \( \text{diam}(I_0)2^{-k}, I_k \) must be contained in \( B(c, r) \) and so has no points of \( C \) in it, contrary to the manner in which the \( I_k \) are defined in which \( f \) is unbounded on \( I_k \cap C \). Therefore, \( c \in C \) as claimed.

Now for \( k \) large enough, and \( x \in C \cap I_k \), the continuity of \( f \) implies \( |f(c) - f(x)| < 1 \) contradicting the manner in which \( I_k \) was chosen since this inequality implies \( f \) is bounded on \( I_k \cap C \). This proves the theorem.

Here is a proof of the extreme value theorem.

**Theorem 12.7.6** Let \( C \) be closed and bounded and let \( f: C \to \mathbb{R} \) be continuous. Then \( f \) achieves its maximum and its minimum on \( C \). This means there exist, \( x_1, x_2 \in C \) such that for all \( x \in C \),

\[
f(x_1) \leq f(x) \leq f(x_2).
\]

**Proof:** Let \( M = \sup \{ f(x) : x \in C \} \). Then by Lemma 12.7.5, \( M \) is a finite number. Is \( f(x_2) = M \) for some \( x_2 \)? if not, you could consider the function,

\[
g(x) \equiv \frac{1}{M - f(x)}
\]

and \( g \) would be a continuous and unbounded function defined on \( C \), contrary to Lemma 12.7.5. Therefore, there exists \( x_2 \in C \) such that \( f(x_2) = M \). A similar argument applies to show the existence of \( x_1 \in C \) such that

\[
f(x_1) = \inf \{ f(x) : x \in C \}.
\]

This proves the theorem.

### 12.7.3 Sequences And Completeness

**Definition 12.7.7** A function whose domain is defined as a set of the form \( \{k, k+1, k+2, \cdots \} \) for \( k \) an integer is known as a sequence. Thus you can consider \( f(k), f(k+1), f(k+2), \) etc. Usually the domain of the sequence is either \( \mathbb{N} \), the natural numbers consisting of \( \{1, 2, 3, \cdots \} \) or the nonnegative integers, \( \{0, 1, 2, 3, \cdots \} \). Also, it is traditional to write \( f_1, f_2, \) etc. instead of \( f(1), f(2), f(3) \) etc. when referring to sequences. In the above context, \( f_k \) is called the first term, \( f_{k+1} \) the second and so forth. It is also common to write the sequence, not as \( f \) but as \( \{ f_i \} \) or just \( \{ f_i \} \) for short. The letter used for the name of the sequence is not important. Thus it is all right to let \( a \) be the name of a sequence or to refer to it as \( \{ a_i \} \). When the sequence has values in \( \mathbb{R}^p \), it is customary to write it in bold face. Thus \( \{ a_i \} \) would refer to a sequence having values in \( \mathbb{R}^p \) for some \( p > 1 \).

**Example 12.7.8** Let \( \{ a_k \} \) be defined by \( a_k \equiv k^2 + 1 \).

This gives a sequence. In fact, \( a_7 = a(7) = 7^2 + 1 = 50 \) just from using the formula for the \( k^{th} \) term of the sequence.

It is nice when sequences come to us in this way from a formula for the \( k^{th} \) term. However, this is often not the case. Sometimes sequences are defined recursively. This happens, when the first several terms of the sequence are given and then a rule is specified which determines \( a_{n+1} \) from knowledge of \( a_1, \cdots, a_n \). This rule which specifies \( a_{n+1} \) from knowledge of \( a_k \) for \( k \leq n \) is known as a recurrence relation.
Example 12.7.9 Let $a_1 = 1$ and $a_2 = 1$. Assuming $a_1, \cdots, a_{n+1}$ are known, $a_{n+2} \equiv a_n + a_{n+1}$.

Thus the first several terms of this sequence, listed in order, are 1, 1, 2, 3, 5, 8, $\cdots$.
This particular sequence is called the Fibonacci sequence and is important in the study of reproducing rabbits.

Example 12.7.10 Let $a_k = (k, \sin (k))$. Thus this sequence has values in $\mathbb{R}^2$.

Definition 12.7.11 Let $\{a_n\}$ be a sequence and let $n_1 < n_2 < n_3, \cdots$ be any strictly increasing list of integers such that $n_1$ is at least as large as the first index used to define the sequence $\{a_n\}$. Then if $b_k \equiv a_{n_k}$, $\{b_k\}$ is called a subsequence of $\{a_n\}$.

For example, suppose $a_n = (n^2 + 1)$. Thus $a_1 = 2$, $a_3 = 10$, etc. If

$n_1 = 1, n_2 = 3, n_3 = 5, \cdots, n_k = 2k - 1$,

then letting $b_k = a_{n_k}$, it follows

$\displaystyle b_k = \left( (2k-1)^2 + 1 \right) = 4k^2 - 4k + 2$.

Definition 12.7.12 A sequence, $\{a_k\}$ is said to converge to $a$ if for every $\varepsilon > 0$ there exists $n_\varepsilon$ such that if $n > n_\varepsilon$ then $|a_n - a| < \varepsilon$. The usual notation for this is $\lim_{n \to \infty} a_n = a$ although it is often written as $a_n \to a$.

The following theorem says the limit, if it exists, is unique.

Theorem 12.7.13 If a sequence, $\{a_n\}$ converges to $a$ and to $b$ then $a = b$.

Proof: There exists $n_\varepsilon$ such that if $n > n_\varepsilon$ then $|a_n - a| < \frac{\varepsilon}{2}$ and if $n > n_\varepsilon$, then $|a_n - b| < \frac{\varepsilon}{2}$. Then pick such an $n$.

$|a - b| < |a - a_n| + |a_n - b| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$.

Since $\varepsilon$ is arbitrary, this proves the theorem.

The following is the definition of a Cauchy sequence in $\mathbb{R}^p$.

Definition 12.7.14 $\{a_n\}$ is a Cauchy sequence if for all $\varepsilon > 0$, there exists $n_\varepsilon$ such that whenever $n, m \geq n_\varepsilon$,

$|a_n - a_m| < \varepsilon$.

A sequence is Cauchy means the terms are “bunching up to each other” as $m, n$ get large.

Theorem 12.7.15 The set of terms in a Cauchy sequence in $\mathbb{R}^p$ is bounded in the sense that for all $n$, $|a_n| < M$ for some $M < \infty$.

Proof: Let $\varepsilon = 1$ in the definition of a Cauchy sequence and let $n > n_1$. Then from the definition,

$|a_n - a_{n_1}| < 1$.

It follows that for all $n > n_1$,

$|a_n| < 1 + |a_{n_1}|$.

Therefore, for all $n$,

$|a_n| \leq 1 + |a_{n_1}| + \sum_{k=1}^{n_1} |a_k|$.

This proves the theorem.
Theorem 12.7.16 If a sequence \( \{a_n\} \) in \( \mathbb{R}^p \) converges, then the sequence is a Cauchy sequence. Also, if some subsequence of a Cauchy sequence converges, then the original sequence converges.

Proof: Let \( \varepsilon > 0 \) be given and suppose \( a_n \to a \). Then from the definition of convergence, there exists \( n_\varepsilon \) such that if \( n > n_\varepsilon \), it follows that

\[
|a_n - a| < \frac{\varepsilon}{2}.
\]

Therefore, if \( m, n \geq n_\varepsilon + 1 \), it follows that

\[
|a_n - a_m| \leq |a_n - a| + |a - a_m| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon
\]

showing that, since \( \varepsilon > 0 \) is arbitrary, \( \{a_n\} \) is a Cauchy sequence. It remains to show the last claim. Suppose then that \( \{a_n\} \) is a Cauchy sequence and \( a = \lim_{k \to \infty} a_{n_k} \) where \( \{a_{n_k}\}_{k=1}^\infty \) is a subsequence. Let \( \varepsilon > 0 \) be given. Then there exists \( K \) such that if \( k, l \geq K \), then \( |a_k - a_l| < \frac{\varepsilon}{2} \). Then if \( k, l > K \), it follows \( n_k > K \) because \( n_1, n_2, n_3, \ldots \) is strictly increasing as the subscript increases. Also, there exists \( K_1 \) such that if \( k > K_1, |a_{n_k} - a| < \frac{\varepsilon}{2} \). Then letting \( n > \max(K, K_1) \), pick \( k > \max(K, K_1) \). Then

\[
|a - a_n| \leq |a - a_{n_k}| + |a_{n_k} - a| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\]

This proves the theorem.

Definition 12.7.17 A set, \( K \) in \( \mathbb{R}^p \) is said to be sequentially compact if every sequence in \( K \) has a subsequence which converges to a point of \( K \).

Theorem 12.7.18 If \( I_0 = \prod_{i=1}^p [a_i, b_i] \) where \( a_i \leq b_i \), then \( I_0 \) is sequentially compact.

Proof: Let \( \{a_i\}_{i=1}^\infty \subseteq I_0 \) and consider all sets of the form \( \prod_{i=1}^p [c_i, d_i] \) where \( [c_i, d_i] \) equals either \( [a_i, \frac{a_i + b_i}{2}] \) or \( [\frac{a_i + b_i}{2}, b_i] \). Thus there are \( 2^p \) of these sets because there are two choices for the \( i \)th slot for \( i = 1, \ldots, p \). Also, if \( x \) and \( y \) are two points in one of these sets,

\[
|x_i - y_i| \leq 2^{-1} |b_i - a_i|.
\]

diam \( (I_0) = \left( \sum_{i=1}^p |b_i - a_i|^2 \right)^{1/2} \),

\[
|x - y| = \left( \sum_{i=1}^p |x_i - y_i|^2 \right)^{1/2}
\leq 2^{-1} \left( \sum_{i=1}^p |b_i - a_i|^2 \right)^{1/2} = 2^{-1} \text{diam}(I_0).
\]

In particular, since \( d \equiv (d_1, \ldots, d_p) \) and \( c \equiv (c_1, \ldots, c_p) \) are two such points,

\[
D_1 \equiv \left( \sum_{i=1}^p |d_i - c_i|^2 \right)^{1/2} \leq 2^{-1} \text{diam}(I_0).
\]

Denote by \( \{J_1, \ldots, J_{2^p}\} \) these sets determined above. Since the union of these sets equals all of \( I_0 \equiv I \), it follows that for some \( J_k \), the sequence, \( \{a_i\} \) is contained in \( J_k \) for infinitely many \( k \). Let that one be called \( I_1 \). Next do for \( I_1 \) what was done for \( I_0 \) to get \( I_2 \subseteq I_1 \) such
that the diameter is half that of $I_1$ and $I_2$ contains $\{a_k\}$ for infinitely many values of $k$. Continue in this way obtaining a nested sequence of intervals, $\{I_k\}$ such that $I_k \supseteq I_{k+1}$, and if $x, y \in I_k$, then $|x - y| \leq 2^{-k} \text{diam}(I_0)$, and $I_n$ contains $\{a_k\}$ for infinitely many values of $k$ for each $n$. Then by the nested interval lemma, there exists $c$ such that $c$ is contained in each $I_k$. Pick $a_{n_1} \in I_1$. Next pick $n_2 > n_1$ such that $a_{n_2} \in I_2$. If $a_{n_1}, \ldots, a_{n_k}$ have been chosen, let $a_{n_{k+1}} \in I_{k+1}$ and $n_{k+1} > n_k$. This can be done because in the construction, $I_n$ contains $\{a_k\}$ for infinitely many $k$. Thus the distance between $a_{n_k}$ and $c$ is no larger than $2^{-k} \text{diam}(I_0)$ and so $\lim_{k \to \infty} a_{n_k} = c \in I_0$. This proves the theorem.

**Theorem 12.7.19** Every Cauchy sequence in $\mathbb{R}^p$ converges.

**Proof:** Let $\{a_k\}$ be a Cauchy sequence. By Theorem 12.7.15 there is some interval, $\prod_{i=1}^{p} [a_i, b_i]$ containing all the terms of $\{a_k\}$. Therefore, by Theorem 12.7.18 a subsequence converges to a point of this interval. By Theorem 12.7.16 the original sequence converges. This proves the theorem.

### 12.7.4 Continuity And The Limit Of A Sequence

Just as in the case of a function of one variable, there is a very useful way of thinking of continuity in terms of limits of sequences found in the following theorem. In words, it says a function is continuous if it takes convergent sequences to convergent sequences whenever possible.

**Theorem 12.7.20** A function $f : D(f) \to \mathbb{R}^n$ is continuous at $x \in D(f)$ if and only if, whenever $x_n \to x$ with $x_n \in D(f)$, it follows $f(x_n) \to f(x)$.

**Proof:** Suppose first that $f$ is continuous at $x$ and let $x_n \to x$. Let $\varepsilon > 0$ be given. By continuity, there exists $\delta > 0$ such that if $|y - x| < \delta$, then $|f(y) - f(x)| < \varepsilon$. However, there exists $n_\delta$ such that if $n \geq n_\delta$, then $|x_n - x| < \delta$ and so for all $n$ this large,

$$|f(x) - f(x_n)| < \varepsilon$$

which shows $f(x_n) \to f(x)$.

Now suppose the condition about taking convergent sequences to convergent sequences holds at $x$. Suppose $f$ fails to be continuous at $x$. Then there exists $\varepsilon > 0$ and $x_n \in D(f)$ such that $|x - x_n| < \frac{1}{n}$, yet

$$|f(x) - f(x_n)| \geq \varepsilon.$$

But this is clearly a contradiction because, although $x_n \to x$, $f(x_n)$ fails to converge to $f(x)$. It follows $f$ must be continuous after all. This proves the theorem.

### 12.8 Exercises

1. Suppose $\{x_n\}$ is a sequence contained in a closed set, $C$ which converges to $x$. Show that $x \in C$. **Hint:** Recall that a set is closed if and only if the complement of the set is open. That is if and only if $\mathbb{R}^n \setminus C$ is open.

2. Show using Problem 1 and Theorem 12.7.18 that every closed and bounded set is sequentially compact. **Hint:** If $C$ is such a set, then $C \subseteq I_0 \equiv \prod_{i=1}^{n} [a_i, b_i]$. Now if $\{x_n\}$ is a sequence in $C$, it must also be a sequence in $I_0$. Apply Problem 1 and Theorem 12.7.18.
3. Prove the extreme value theorem, a continuous function achieves its maximum and minimum on any closed and bounded set, $C$, using the result of Problem 2. **Hint:** Suppose $\lambda = \sup \{ f(x) : x \in C \}$ . Then there exists $\{x_n\} \subseteq C$ such that $f(x_n) \to \lambda$. Now select a convergent subsequence using Problem 2. Do the same for the minimum.

4. Let $C$ be a closed and bounded set and suppose $f : C \to \mathbb{R}^n$ is continuous. Show that $f$ must also be uniformly continuous. This means: For every $\varepsilon > 0$ there exists $\delta > 0$ such that whenever $x, y \in C$ and $|x - y| < \delta$, it follows $|f(x) - f(y)| < \varepsilon$. This is a good time to review the definition of continuity so you will see the difference. **Hint:** Suppose it is not so. Then there exists $\varepsilon > 0$ and $\{x_k\}$ and $\{y_k\}$ such that $|x_k - y_k| < \frac{1}{k}$ but $|f(x_k) - f(y_k)| \geq \varepsilon$. Now use Problem 2 to obtain a convergent subsequence.

5. Suppose every Cauchy sequence converges in $\mathbb{R}$. Show this implies the least upper bound axiom which is the usual way to state completeness for $\mathbb{R}$. Explain why the convergence of Cauchy sequences is equivalent to every nonempty set which is bounded above has a least upper bound in $\mathbb{R}$.

6. From Problem 2 every closed and bounded set is sequentially compact. Are these the only sets which are sequentially compact? Explain.

7. A set whose elements are open sets, $\mathcal{C}$ is called an open cover of $H$ if $\cup \mathcal{C} \supseteq H$. In other words, $\mathcal{C}$ is an open cover of $H$ if every point of $H$ is in at least one set of $\mathcal{C}$. Show that if $\mathcal{C}$ is an open cover of a closed and bounded set $H$ then there exists $\delta > 0$ such that whenever $x \in H$, $B(x, \delta)$ is contained in some set of $\mathcal{C}$. This number, $\delta$ is called a Lebesgue number. **Hint:** If there is no Lebesgue number for $H$, let $H \subseteq I = \prod_{i=1}^n [a_i, b_i]$. Use the process of chopping the intervals in half to get a sequence of nested intervals, $I_k$ contained in $I$ where $\text{diam}(I_k) \leq 2^{-k} \text{diam}(I)$ and there is no Lebesgue number for the open cover on $H_k \equiv H \cap I_k$. Now use the nested interval theorem to get $c$ in all these $H_k$. For some $r > 0$ it follows $B(c, r)$ is contained in some open set of $U$. But for large $k$, it must be that $H_k \subseteq B(c, r)$ which contradicts the construction. You fill in the details.

8. A set is compact if for every open cover of the set, there exists a finite subset of the open cover which also covers the set. Show every closed and bounded set in $\mathbb{R}^p$ is compact. Next show that if a set in $\mathbb{R}^p$ is compact, then it must be closed and bounded. This is called the Heine Borel theorem.

9. Suppose $S$ is a nonempty set in $\mathbb{R}^p$. Define

$$ \text{dist}(x, S) \equiv \inf \{|x - y| : y \in S\}.$$ 

Show that

$$|\text{dist}(x, S) - \text{dist}(y, S)| \leq |x - y|.$$ 

**Hint:** Suppose $\text{dist}(x, S) < \text{dist}(y, S)$. If these are equal there is nothing to show. Explain why there exists $z \in S$ such that $|x - z| < \text{dist}(x, S) + \varepsilon$. Now explain why

$$|\text{dist}(x, S) - \text{dist}(y, S)| = \text{dist}(y, S) - \text{dist}(x, S) \leq |y - z| - (|x - z| - \varepsilon)$$

Now use the triangle inequality and observe that $\varepsilon$ is arbitrary.

10. Suppose $H$ is a closed set and $H \subseteq U \subseteq \mathbb{R}^p$, an open set. Show there exists a continuous function defined on $\mathbb{R}^p$, $f$ such that $f(\mathbb{R}^p) \subseteq [0, 1]$, $f(x) = 0$ if $x \notin U$ and
\[ f(x) = 1 \text{ if } x \in H. \textbf{Hint:} \text{ Try something like} \]
\[
\frac{\text{dist}(x, U^C)}{\text{dist}(x, U^C) + \text{dist}(x, H)},
\]
where \( U^C \equiv \mathbb{R}^p \setminus U \), a closed set. You need to explain why the denominator is never equal to zero. The rest is supplied by Problem 9. This is a special case of a major theorem called Urysohn’s lemma.
Vector Valued Functions Of One Variable

13.0.1 Outcomes

1. Identify a curve given its parameterization.
2. Determine combinations of vector functions such as sums, vector products, and scalar products.
3. Define limit, derivative, and integral for vector functions.
4. Evaluate limits, derivatives and integrals of vector functions.
5. Find the line tangent to a curve at a given point.
6. Recall, derive and apply rules to combinations of vector functions for the following:
   (a) limits
   (b) differentiation
   (c) integration
7. Describe what is meant by arc length.
8. Evaluate the arc length of a curve.
9. Evaluate the work done by a varying force over a curved path.

13.1 Limits Of A Vector Valued Function Of One Variable

The above discussion considered expressions like

\[
f(t_0 + h) - f(t_0)
\]

and determined what they get close to as \( h \) gets small. In other words it is desired to consider

\[
\lim_{h \to 0} \frac{f(t_0 + h) - f(t_0)}{h}
\]
Specializing to functions of one variable, one can give a meaning to
\[ \lim_{s \to t^+} f(s), \lim_{s \to t^-} f(s), \lim_{s \to \infty} f(s), \]
and
\[ \lim_{s \to -\infty} f(s). \]

**Definition 13.1.1** In the case where \( D(f) \) is only assumed to satisfy \( D(f) \supseteq (t, t + r) \),
\[ \lim_{s \to t^+} f(s) = L \]
if and only if for all \( \varepsilon > 0 \) there exists \( \delta > 0 \) such that if
\[ 0 < s - t < \delta, \]
then
\[ |f(s) - L| < \varepsilon. \]

In the case where \( D(f) \) is only assumed to satisfy \( D(f) \supseteq (t - r, t) \),
\[ \lim_{s \to t^-} f(s) = L \]
if and only if for all \( \varepsilon > 0 \) there exists \( \delta > 0 \) such that if
\[ 0 < t - s < \delta, \]
then
\[ |f(s) - L| < \varepsilon. \]

One can also consider limits as a variable “approaches” infinity. Of course nothing is “close”
to infinity and so this requires a slightly different definition.
\[ \lim_{t \to \infty} f(t) = L \]
if for every \( \varepsilon > 0 \) there exists \( l \) such that whenever \( t > l \),
\[ |f(t) - L| < \varepsilon \] (13.1)
and
\[ \lim_{t \to -\infty} f(t) = L \]
if for every \( \varepsilon > 0 \) there exists \( l \) such that whenever \( t < l \), (13.1) holds.

Note that in all of this the definitions are identical to the case of scalar valued functions. The only difference is that here \( |\cdot| \) refers to the norm or length in \( \mathbb{R}^p \) where maybe \( p > 1 \).

**Example 13.1.2** Let \( f(t) = (\cos t, \sin t, t^2 + 1, \ln(t)) \). Find \( \lim_{t \to \pi/2} f(t) \).

Use Theorem 12.4.7 on Page 215 and the continuity of the functions to write this limit equals
\[ \left( \lim_{t \to \pi/2} \cos t, \lim_{t \to \pi/2} \sin t, \lim_{t \to \pi/2} (t^2 + 1), \lim_{t \to \pi/2} \ln(t) \right) \]
\[ = \left( 0, 1, \ln\left(\frac{\pi^2}{4} + 1\right), \ln\left(\frac{\pi}{2}\right) \right). \]

**Example 13.1.3** Let \( f(t) = (\sin t, t^2, t + 1) \). Find \( \lim_{t \to 0} f(t) \).

Recall that \( \lim_{t \to 0} \sin t = 1 \). Then from Theorem 12.4.7 on Page 215, \( \lim_{t \to 0} f(t) = (1, 0, 1) \).
13.2 The Derivative And Integral

The following definition is on the derivative and integral of a vector valued function of one variable.

**Definition 13.2.1** The derivative of a function, \( f'(t) \), is defined as the following limit whenever the limit exists. If the limit does not exist, then neither does \( f'(t) \).

\[
\lim_{h \to 0} \frac{f(t+h) - f(t)}{h} \equiv f'(t)
\]

The function of \( h \) on the left is called the difference quotient just as it was for a scalar valued function. If \( f(t) = (f_1(t), \ldots, f_p(t)) \) and \( \int_a^b f_i(t) \, dt \) exists for each \( i = 1, \ldots, p \), then \( \int_a^b f(t) \, dt \) is defined as the vector,

\[
\left( \int_a^b f_1(t) \, dt, \ldots, \int_a^b f_p(t) \, dt \right).
\]

This is what is meant by saying \( f \in R([a,b]) \).

This is exactly like the definition for a scalar valued function. As before,

\[
f'(x) = \lim_{y \to x} \frac{f(y) - f(x)}{y - x}.
\]

As in the case of a scalar valued function, differentiability implies continuity but not the other way around.

**Theorem 13.2.2** If \( f'(t) \) exists, then \( f \) is continuous at \( t \).

**Proof:** Suppose \( \varepsilon > 0 \) is given and choose \( \delta_1 > 0 \) such that if \( |h| < \delta_1 \),

\[
\left| \frac{f(t+h) - f(t)}{h} - f'(t) \right| < 1.
\]

then for such \( h \), the triangle inequality implies

\[
|f(t+h) - f(t)| < |h| + |f'(t)| \cdot |h|.
\]

Now letting \( \delta < \min \left( \delta_1, \frac{\varepsilon}{|f'(x)|} \right) \) it follows if \( |h| < \delta \), then

\[
|f(t+h) - f(t)| < \varepsilon.
\]

Letting \( y = h + t \), this shows that if \( |y - t| < \delta \),

\[
|f(y) - f(t)| < \varepsilon
\]

which proves \( f \) is continuous at \( t \). This proves the theorem.

As in the scalar case, there is a fundamental theorem of calculus.

**Theorem 13.2.3** If \( f \in R([a,b]) \) and if \( f \) is continuous at \( t \in (a,b) \), then

\[
\frac{d}{dt} \left( \int_a^t f(s) \, ds \right) = f(t).
\]
\textbf{Proof:} Say $f(t) = (f_1(t), \cdots, f_p(t))$. Then it follows
\[
\frac{1}{h} \int_a^{t+h} f(s) \, ds - \frac{1}{h} \int_a^t f(s) \, ds = \left( \frac{1}{h} \int_t^{t+h} f_1(s) \, ds, \cdots, \frac{1}{h} \int_t^{t+h} f_p(s) \, ds \right)
\]
and $\lim_{h \to 0} \frac{1}{h} \int_t^{t+h} f_i(s) \, ds = f_i(t)$ for each $i = 1, \cdots, p$ from the fundamental theorem of calculus for scalar valued functions. Therefore,
\[
\lim_{h \to 0} \frac{1}{h} \int_a^{t+h} f(s) \, ds - \frac{1}{h} \int_a^t f(s) \, ds = (f_1(t), \cdots, f_p(t)) = f(t)
\]
and this proves the claim.

\textbf{Example 13.2.4} Let $f(x) = c$ where $c$ is a constant. Find $f'(x)$.

The difference quotient,
\[
\frac{f(x+h) - f(x)}{h} = \frac{c - c}{h} = 0
\]
Therefore,
\[
\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = 0 = 0
\]

\textbf{Example 13.2.5} Let $f(t) = (at, bt)$ where $a, b$ are constants. Find $f'(t)$.

From the above discussion this derivative is just the vector valued functions whose components consist of the derivatives of the components of $f$. Thus $f'(t) = (a, b)$.

\section*{13.2.1 Geometric And Physical Significance Of The Derivative}

Suppose $r$ is a vector valued function of a parameter, $t$ not necessarily time and consider the following picture of the points traced out by $r$.

In this picture there are unit vectors in the direction of the vector from $r(t)$ to $r(t+h)$. You can see that it is reasonable to suppose these unit vectors, if they converge, converge to a unit vector, $T$ which is tangent to the curve at the point $r(t)$. Now each of these unit vectors is of the form $r(t+h) - r(t)$. Therefore,
\[
r'(t) = \lim_{h \to 0} \frac{r(t+h) - r(t)}{h} = \lim_{h \to 0} \frac{|r(t+h) - r(t)|}{|r(t+h) - r(t)|} \cdot \frac{r(t+h) - r(t)}{|r(t+h) - r(t)|} = \lim_{h \to 0} \frac{|r(t+h) - r(t)|}{h} T_h = |r'(t)| T.
\]
In the case that \( t \) is time, the expression \(|r(t + h) - r(t)|\) is a good approximation for the distance traveled by the object on the time interval \([t, t + h]\). The real distance would be the length of the curve joining the two points but if \( h \) is very small, this is essentially equal to \(|r(t + h) - r(t)|\) as suggested by the picture below.

Therefore, \[
\frac{|r(t + h) - r(t)|}{h}
\]
gives for small \( h \), the approximate distance travelled on the time interval \([t, t + h]\) divided by the length of time, \( h \). Therefore, this expression is really the average speed of the object on this small time interval and so the limit as \( h \to 0 \), deserves to be called the instantaneous speed of the object. Thus \( r'(t) \) represents the speed times a unit direction vector, \( T \) which defines the direction in which the object is moving. Thus \( r'(t) \) is the velocity of the object. This is the physical significance of the derivative when \( t \) is time.

Example 13.2.6 Let \( r(t) = (\sin t, t^2, t + 1) \) for \( t \in [0, 5] \). Find a tangent line to the curve parameterized by \( r \) at the point \( r(2) \).

From the above discussion, a direction vector has the same direction as \( r'(2) \). Therefore, it suffices to simply use \( r'(2) \) as a direction vector for the line. \( r'(2) = (\cos 2, 4, 1) \). Therefore, a parametric equation for the tangent line is
\[
(s \sin 2, 4, 3) + t(\cos 2, 4, 1) = (x, y, z).
\]
Example 13.2.7 Let \( \mathbf{r} (t) = (\sin t, t^2, t + 1) \) for \( t \in [0, 5] \). Find the velocity vector when \( t = 1 \).

From the above discussion, this is simply \( \mathbf{r}' (1) = (\cos 1, 2, 1) \).

13.2.2 Differentiation Rules

There are rules which relate the derivative to the various operations done with vectors such as the dot product, the cross product, and vector addition and scalar multiplication.

**Theorem 13.2.8** Let \( a, b \in \mathbb{R} \) and suppose \( \mathbf{f}' (t) \) and \( \mathbf{g}' (t) \) exist. Then the following formulas are obtained.

\[
(af + bg)' (t) = af' (t) + bg' (t),
\]

(13.3)

\[
(f \cdot g)' (t) = f' (t) \cdot g (t) + f (t) \cdot g' (t)
\]

(13.4)

If \( \mathbf{f}, \mathbf{g} \) have values in \( \mathbb{R}^3 \), then

\[
(f \times g)' (t) = f(t) \times g'(t) + f'(t) \times g(t)
\]

(13.5)

The formulas, (13.4), and (13.5) are referred to as the product rule.

**Proof:** The first formula is left for you to prove. Consider the second, (13.4).

\[
\lim_{h \to 0} \frac{\mathbf{f} \cdot \mathbf{g}(t + h) - \mathbf{f} \cdot \mathbf{g}(t)}{h}
\]

\[
= \lim_{h \to 0} \frac{\mathbf{f}(t + h) \cdot \mathbf{g}(t + h) - \mathbf{f}(t + h) \cdot \mathbf{g}(t) + \mathbf{f}(t + h) \cdot \mathbf{g}(t) - \mathbf{f}(t) \cdot \mathbf{g}(t)}{h}
\]

\[
= \lim_{h \to 0} \left( \mathbf{f}(t + h) \cdot \frac{\mathbf{g}(t + h) - \mathbf{g}(t)}{h} + \frac{\mathbf{f}(t + h) - \mathbf{f}(t)}{h} \cdot \mathbf{g}(t) \right)
\]

\[
= \lim_{h \to 0} \sum_{k=1}^{n} f_k (t + h) \frac{g_k (t + h) - g_k (t)}{h} + \sum_{k=1}^{n} \frac{f_k (t + h) - f_k (t)}{h} g_k (t)
\]

\[
= \sum_{k=1}^{n} f_k (t) g'_k (t) + \sum_{k=1}^{n} f'_k (t) g_k (t)
\]

\[
= \mathbf{f}' (t) \cdot \mathbf{g} (t) + \mathbf{f} (t) \cdot \mathbf{g}' (t).
\]

Formula (13.5) is left as an exercise which follows from the product rule and the definition of the cross product in terms of components given on Page 143.

**Example 13.2.9** Let

\[
\mathbf{r} (t) = (t^2, \sin t, \cos t)
\]

and let \( \mathbf{p} (t) = (t, \ln (t + 1), 2t) \). Find \( (\mathbf{r} \times \mathbf{p})(t)' \).

From (13.5) this equals \((2t, \cos t, -\sin t) \times (t, \ln (t + 1), 2t) + (t^2, \sin t, \cos t) \times \left(1, \frac{1}{t+1}, 2\right)\).

**Example 13.2.10** Let \( \mathbf{r} (t) = (t^2, \sin t, \cos t) \) Find \( \int_{0}^{\pi} \mathbf{r} (t) \, dt \).

This equals \((\int_{0}^{\pi} t^2 \, dt, \int_{0}^{\pi} \sin t \, dt, \int_{0}^{\pi} \cos t \, dt) = (\frac{1}{3} \pi^3, 2, 0)\).
Example 13.2.11 An object has position \( \mathbf{r}(t) = \left( t^2, \frac{1}{1+t}, \sqrt{t^2 + 2} \right) \) kilometers where \( t \) is given in hours. Find the velocity of the object in kilometers per hour when \( t = 1 \).

Recall the velocity at time \( t \) was \( \mathbf{r}'(t) \). Therefore, find \( \mathbf{r}'(t) \) and plug in \( t = 1 \) to find the velocity.

\[
\mathbf{r}'(t) = \left( 3t^2, \frac{1}{1+t} - t, \frac{1}{2}(t^2 + 2)^{-1/2} \right) \\
= \left( 3t^2, \frac{1}{1+t} - t, \frac{1}{\sqrt{(t^2 + 2)}} \right)
\]

When \( t = 1 \), the velocity is

\[
\mathbf{r}'(1) = \left( 3, -\frac{3}{2}, \frac{1}{\sqrt{3}} \right) \text{ kilometers per hour.}
\]

Obviously, this can be continued. That is, you can consider the possibility of taking the derivative of the derivative and then the derivative of that and so forth. The main thing to consider about this is the notation and it is exactly like it was in the case of a scalar valued function presented earlier. Thus \( \mathbf{r}''(t) \) denotes the second derivative.

When you are given a vector valued function of one variable, sometimes it is possible to give a simple description of the curve which results. Usually it is not possible to do this!

Example 13.2.12 Describe the curve which results from the vector valued function, \( \mathbf{r}(t) = (\cos 2t, \sin 2t, t) \) where \( t \in \mathbb{R} \).

The first two components indicate that for \( \mathbf{r}(t) = (x(t), y(t), z(t)) \), the pair, \( (x(t), y(t)) \) traces out a circle. While it is doing so, \( z(t) \) is moving at a steady rate in the positive direction. Therefore, the curve which results is a cork skrew shaped thing called a helix.

As an application of the theorems for differentiating curves, here is an interesting application. It is also a situation where the curve can be identified as something familiar.

Example 13.2.13 Sound waves have the angle of incidence equal to the angle of reflection. Suppose you are in a large room and you make a sound. The sound waves spread out and you would expect your sound to be inaudible very far away. But what if the room were shaped so that the sound is reflected off the wall toward a single point, possibly far away from you? Then you might have the interesting phenomenon of someone far away hearing what you said quite clearly. How should the room be designed?

Suppose you are located at the point \( \mathbf{P}_0 \) and the point where your sound is to be reflected is \( \mathbf{P}_1 \). Consider a plane which contains the two points and let \( \mathbf{r}(t) \) denote a parameterization of the intersection of this plane with the walls of the room. Then the condition that the angle of reflection equals the angle of incidence reduces to saying the angle between \( \mathbf{P}_0 - \mathbf{r}(t) \) and \( -\mathbf{r}'(t) \) equals the angle between \( \mathbf{P}_1 - \mathbf{r}(t) \) and \( \mathbf{r}'(t) \).

Draw a picture to see this. Therefore,

\[
\frac{(\mathbf{P}_0 - \mathbf{r}(t)) \cdot (-\mathbf{r}'(t))}{|\mathbf{P}_0 - \mathbf{r}(t)||\mathbf{r}'(t)|} = \frac{(\mathbf{P}_1 - \mathbf{r}(t)) \cdot (\mathbf{r}'(t))}{|\mathbf{P}_1 - \mathbf{r}(t)||\mathbf{r}'(t)|}.
\]

This reduces to

\[
\frac{(\mathbf{r}(t) - \mathbf{P}_0) \cdot (-\mathbf{r}'(t))}{|\mathbf{r}(t) - \mathbf{P}_0|} = \frac{(\mathbf{r}(t) - \mathbf{P}_1) \cdot (\mathbf{r}'(t))}{|\mathbf{r}(t) - \mathbf{P}_1|} \quad (13.6)
\]
Now
\[
\frac{(\mathbf{r}(t) - \mathbf{P}_1) \cdot (\mathbf{r}'(t))}{|\mathbf{r}(t) - \mathbf{P}_1|} = \frac{d}{dt} |\mathbf{r}(t) - \mathbf{P}_1|
\]
and a similar formula holds for \( \mathbf{P}_1 \) replaced with \( \mathbf{P}_0 \). This is because
\[
|\mathbf{r}(t) - \mathbf{P}_1| = \sqrt{(\mathbf{r}(t) - \mathbf{P}_1) \cdot (\mathbf{r}(t) - \mathbf{P}_1)}
\]
and so using the chain rule and product rule,
\[
\frac{d}{dt} |\mathbf{r}(t) - \mathbf{P}_1| = \frac{1}{2} ((\mathbf{r}(t) - \mathbf{P}_1) \cdot (\mathbf{r}(t) - \mathbf{P}_1))^{-1/2} 2 ((\mathbf{r}(t) - \mathbf{P}_1) \cdot \mathbf{r}'(t))
\]
\[
= \frac{(\mathbf{r}(t) - \mathbf{P}_1) \cdot (\mathbf{r}'(t))}{|\mathbf{r}(t) - \mathbf{P}_1|}.
\]
Therefore, from (13.6),
\[
\frac{d}{dt} (|\mathbf{r}(t) - \mathbf{P}_1|) + \frac{d}{dt} (|\mathbf{r}(t) - \mathbf{P}_0|) = 0
\]
showing that \( |\mathbf{r}(t) - \mathbf{P}_1| + |\mathbf{r}(t) - \mathbf{P}_0| = C \) for some constant, \( C \). This implies the curve of intersection of the plane with the room is an ellipse having \( \mathbf{P}_0 \) and \( \mathbf{P}_1 \) as the foci.

13.2.3 Leibniz’s Notation

Leibniz’s notation also generalizes routinely. For example, \( \frac{d\mathbf{r}}{dt} = \mathbf{y}'(t) \) with other similar notations holding.

13.3 Product Rule For Matrices

Here is the concept of the product rule extended to matrix multiplication.

**Definition 13.3.1** Let \( A(t) \) be an \( m \times n \) matrix. Say \( A(t) = (A_{ij}(t)) \). Suppose also that \( A_{ij}(t) \) is a differentiable function for all \( i, j \). Then define \( A'(t) \equiv (A'_{ij}(t)) \). That is, \( A'(t) \) is the matrix which consists of replacing each entry by its derivative. Such an \( m \times n \) matrix in which the entries are differentiable functions is called a differentiable matrix.

The next lemma is just a version of the product rule.

**Lemma 13.3.2** Let \( A(t) \) be an \( m \times n \) matrix and let \( B(t) \) be an \( n \times p \) matrix with the property that all the entries of these matrices are differentiable functions. Then
\[
(A(t) B(t))' = A'(t) B(t) + A(t) B'(t).
\]

**Proof:** \( (A(t) B(t))' = (C'_{ij}(t)) \) where \( C_{ij}(t) = A_{ik}(t) B_{kj}(t) \) and the repeated index summation convention is being used. Therefore,
\[
C'_{ij}(t) = A'_{ik}(t) B_{kj}(t) + A_{ik}(t) B'_{kj}(t)
\]
\[
= (A'(t) B(t))_{ij} + (A(t) B'(t))_{ij}
\]
\[
= (A'(t) B(t) + A(t) B'(t))_{ij}
\]
Therefore, the \( ij^{th} \) entry of \( A(t) B(t) \) equals the \( ij^{th} \) entry of \( A'(t) B(t) + A(t) B'(t) \) and this proves the lemma.
13.4 Moving Coordinate Systems

Let \( i(t), j(t), k(t) \) be a right handed\(^1\) orthonormal basis of vectors for each \( t \). It is assumed these vectors are \( C^1 \) functions of \( t \). Letting the positive \( x \) axis extend in the direction of \( i(t) \), the positive \( y \) axis extend in the direction of \( j(t) \), and the positive \( z \) axis extend in the direction of \( k(t) \), yields a moving coordinate system. Now let \( u = (u_1, u_2, u_3) \in \mathbb{R}^3 \) and let \( t_0 \) be some reference time. For example you could let \( t_0 = 0 \). Then define the components of \( u \) with respect to these vectors, \( i, j, k \) at time \( t_0 \) as

\[
u = u_1 i(t_0) + u_2 j(t_0) + u_3 k(t_0) .\]

Let \( u(t) \) be defined as the vector which has the same components with respect to \( i, j, k \) but at time \( t \). Thus

\[
u(t) = u_1 i(t) + u_2 j(t) + u_3 k(t) .\]

and the vector has changed although the components have not.

For example, this is exactly the situation in the case of apparently fixed basis vectors on the earth if \( u \) is a position vector from the given spot on the earth’s surface to a point regarded as fixed with the earth due to its keeping the same coordinates relative to coordinate axes which are fixed with the earth.

Now define a linear transformation \( Q(t) \) mapping \( \mathbb{R}^3 \) to \( \mathbb{R}^3 \) by

\[
Q(t) u = u_1 i(t) + u_2 j(t) + u_3 k(t)
\]

where

\[
u = u_1 i(t_0) + u_2 j(t_0) + u_3 k(t_0) .\]

Thus letting \( v, u \in \mathbb{R}^3 \) be vectors and \( \alpha, \beta \), scalars,

\[
Q(t) (\alpha u + \beta v) \equiv (\alpha u_1 + \beta v_1) i(t) + (\alpha u_2 + \beta v_2) j(t) + (\alpha u_3 + \beta v_3) k(t)
\]

and

\[
= (\alpha u_1 i(t) + \alpha u_2 j(t) + \alpha u_3 k(t)) + (\beta v_1 i(t) + \beta v_2 j(t) + \beta v_3 k(t)) = \alpha (u_1 i(t) + u_2 j(t) + u_3 k(t)) + \beta (v_1 i(t) + v_2 j(t) + v_3 k(t)) = \alpha Q(t) u + \beta Q(t) v
\]

showing that \( Q(t) \) is a linear transformation. Also, \( Q(t) \) preserves all distances because, since the vectors, \( i(t), j(t), k(t) \) form an orthonormal set,

\[
|Q(t) u| = \left( \sum_{i=1}^{3} (u_i')^2 \right)^{1/2} = |u| .
\]

For simplicity, let \( i(t) = e_1(t), j(t) = e_2(t), k(t) = e_3(t) \) and \( i(t_0) = e_1(t_0), j(t_0) = e_2(t_0), k(t_0) = e_3(t_0) \). Then using the repeated index summation convention,

\[
u(t) = u_j e_j(t) = u_j e_j(t) \cdot e_i(t_0) e_i(t_0)
\]

and so with respect to the basis, \( i(t_0) = e_1(t_0), j(t_0) = e_2(t_0), k(t_0) = e_3(t_0) \), the matrix of \( Q(t) \) is

\[
Q_{ij}(t) = e_i(t_0) \cdot e_j(t)
\]

Recall this means you take a vector, \( u \in \mathbb{R}^3 \) which is a list of the components of \( u \) with respect to \( i(t_0), j(t_0), k(t_0) \) and when you multiply by \( Q(t) \) you get the components of \( u(t) \) with respect to \( i(t), j(t), k(t) \). I will refer to this matrix as \( Q(t) \) to save notation.

\(^1\)Recall that right handed implies \( i \times j = k \).
Lemma 13.4.1 Suppose \( Q(t) \) is a real, differentiable \( n \times n \) matrix which preserves distances. Then \( Q(t)Q(t)^T = Q(t)^TQ(t) = I \). Also, if \( u(t) \equiv Q(t)u \), then there exists a vector, \( \Omega(t) \) such that

\[
u'(t) = \Omega(t) \times u(t).
\]

**Proof:** Recall that \( (z \cdot w) = \frac{1}{2} \left( |z + w|^2 - |z - w|^2 \right) \). Therefore,

\[
(Q(t)u \cdot Q(t)w) = \frac{1}{4} \left( |Q(t)(u + w)|^2 - |Q(t)(u - w)|^2 \right)
\]

\[
= \frac{1}{4} \left( |u + w|^2 - |u - w|^2 \right)
\]

\[
= (u \cdot w).
\]

This implies

\[
(Q(t)^TQ(t)u \cdot w) = (u \cdot w)
\]

for all \( u, w \). Therefore, \( Q(t)^TQ(t)u = u \) and so \( Q(t)^TQ(t) = Q(t)Q(t)^T = I \). This proves the first part of the lemma.

It follows from the product rule, Lemma 13.3.2 that

\[
Q'(t)Q(t)^T + Q(t)Q'(t)^T = 0
\]

and so

\[
Q'(t)Q(t)^T = - \left( Q(t)Q(t)^T \right)^T.
\]  \hspace{1cm} (13.7)

From the definition, \( Q(t)u = u(t) \),

\[
u'(t) = Q'(t)u = \underbrace{Q'(t)Q(t)^T}_{\equiv u}u(t).
\]

Then writing the matrix of \( Q'(t)Q(t)^T \) with respect to \( i(t_0), j(t_0), k(t_0) \), it follows from (13.7) that the matrix of \( Q'(t)Q(t)^T \) is of the form

\[
\begin{pmatrix}
0 & -\omega_3(t) & \omega_2(t) \\
\omega_3(t) & 0 & -\omega_1(t) \\
-\omega_2(t) & \omega_1(t) & 0
\end{pmatrix}
\]

for some time dependent scalars, \( \omega_i \). Therefore,

\[
\begin{pmatrix}
u_1 \\
u_2 \\
u_3
\end{pmatrix}
'= \begin{pmatrix}
0 & -\omega_3(t) & \omega_2(t) \\
\omega_3(t) & 0 & -\omega_1(t) \\
-\omega_2(t) & \omega_1(t) & 0
\end{pmatrix}
\begin{pmatrix}
u_1 \\
u_2 \\
u_3
\end{pmatrix}
\]

\[
= \begin{pmatrix}
w_2(t)u_3(t) - w_3(t)u_2(t) & w_3(t)u_1(t) - w_1(t)u_3(t) & w_1(t)u_2(t) - w_2(t)u_1(t)
\end{pmatrix}
\]

where the \( u_i \) are the components of the vector \( u(t) \) in terms of the fixed vectors \( i(t_0), j(t_0), k(t_0) \). Therefore,

\[
u'(t) = \Omega(t) \times u(t) = Q'(t)Q(t)^Tu(t)
\]  \hspace{1cm} (13.8)

where

\[
\Omega(t) = \omega_1(t)i(t_0) + \omega_2(t)j(t_0) + \omega_3(t)k(t_0).
\]
because
\[
\mathbf{\Omega}(t) \times \mathbf{u}(t) \equiv \begin{vmatrix} \mathbf{i}(t_0) & \mathbf{j}(t_0) & \mathbf{k}(t_0) \\ w_1 & w_2 & w_3 \\ u_1 & u_2 & u_3 \end{vmatrix} = \begin{vmatrix} \mathbf{i}(t_0)(w_2u_3 - w_3u_2) + \mathbf{j}(t_0)(w_3u_1 - w_1u_3) + \mathbf{k}(t_0)(w_1u_2 - w_2u_1) \\
\end{vmatrix}_{i(t_0)w(t_0)u(t_0)}
\]

This proves the lemma and yields the existence part of the following theorem.

**Theorem 13.4.2** Let \( \mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t) \) be as described. Then there exists a unique vector \( \mathbf{\Omega}(t) \) such that if \( \mathbf{u}(t) \) is a vector whose components are constant with respect to \( \mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t) \), then

\[
\mathbf{u}'(t) = \mathbf{\Omega}(t) \times \mathbf{u}(t).
\]

**Proof:** It only remains to prove uniqueness. Suppose \( \mathbf{\Omega}_1 \) also works. Then

\[
\mathbf{u}(t) = \mathbf{Q}(t) \mathbf{u}
\]

and so

\[
\mathbf{u}'(t) = \mathbf{Q}'(t) \mathbf{u}
\]

for all \( \mathbf{u} \). Therefore,

\[
(\mathbf{\Omega} - \mathbf{\Omega}_1) \times \mathbf{Q}(t) \mathbf{u} = \mathbf{0}
\]

for all \( \mathbf{u} \) and since \( \mathbf{Q}(t) \) is one to one and onto, this implies \( (\mathbf{\Omega} - \mathbf{\Omega}_1) \times \mathbf{w} = \mathbf{0} \) for all \( \mathbf{w} \) and thus \( \mathbf{\Omega} - \mathbf{\Omega}_1 = \mathbf{0} \). This proves the theorem.

**Definition 13.4.3** A **rigid body** in \( \mathbb{R}^3 \) is one with the property that there exists a moving coordinate system with the property that for an observer on the rigid body, the vectors, \( \mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t) \) are constant. More generally, a vector \( \mathbf{u}(t) \) is said to be fixed with the body if to a person on the body, the vector appears to have the same magnitude and same direction independent of \( t \). Thus \( \mathbf{u}(t) \) is fixed with the body if \( \mathbf{u}(t) = u_1\mathbf{i}(t) + u_2\mathbf{j}(t) + u_3\mathbf{k}(t) \).

The following comes from the above discussion.

**Theorem 13.4.4** Let \( B(t) \) be the set of points in three dimensions occupied by a rigid body. Then there exists a vector \( \mathbf{\Omega}(t) \) such that whenever \( \mathbf{u}(t) \) is fixed with the rigid body,

\[
\mathbf{u}'(t) = \mathbf{\Omega}(t) \times \mathbf{u}(t).
\]

### 13.5 Exercises

1. Find the following limits if possible
   
   (a) \( \lim_{x \to x_0^+} \left( \frac{|x|}{x}, \sin x / x, \cos x \right) \)
   
   (b) \( \lim_{x \to 0^+} \left( \frac{x}{x}, \sec x, e^x \right) \)
   
   (c) \( \lim_{x \to 4} \left( \frac{x^2 - 16}{x+4}, x + 7, \frac{\tan 4x}{4x} \right) \)
   
   (d) \( \lim_{x \to \infty} \left( \frac{x}{1+x^2}, \frac{x^2}{1+x^2}, \frac{\sin x^2}{x} \right) \)

2. Find \( \lim_{x \to 2} \left( \frac{x^2 - 4}{x^2 + 2x}, x^2 + 2x - 1, \frac{x^2 - 4}{x - 2} \right) \).
3. Prove from the definition that \( \lim_{x \to a} (\sqrt[3]{x}, x + 1) = (\sqrt[3]{a}, a + 1) \) for all \( a \in \mathbb{R} \). **Hint:** You might want to use the formula for the difference of two cubes, 

\[
a^3 - b^3 = (a - b)(a^2 + ab + b^2).
\]

4. Let \( \mathbf{r}(t) = (4 + (t-1)^2, \sqrt{t^2+1}(t-1)^3, \frac{(t-1)^3}{t^2}) \) describe the position of an object in \( \mathbb{R}^3 \) as a function of \( t \) where \( t \) is measured in seconds and \( \mathbf{r}(t) \) is measured in meters. Is the velocity of this object ever equal to zero? If so, find the value of \( t \) at which this occurs and the point in \( \mathbb{R}^3 \) at which the velocity is zero.

5. Let \( \mathbf{r}(t) = (\sin 2t, t^2, 2t + 1) \) for \( t \in [0, 4] \). Find a tangent line to the curve parameterized by \( \mathbf{r} \) at the point \( \mathbf{r}(2) \).

6. Let \( \mathbf{r}(t) = (t, \sin t^2, t + 1) \) for \( t \in [0, 5] \). Find a tangent line to the curve parameterized by \( \mathbf{r} \) at the point \( \mathbf{r}(2) \).

7. Let \( \mathbf{r}(t) = (\sin t, t^2, \cos (t^2)) \) for \( t \in [0, 5] \). Find a tangent line to the curve parameterized by \( \mathbf{r} \) at the point \( \mathbf{r}(2) \).

8. Let \( \mathbf{r}(t) = (\sin t, \cos (t^2), t + 1) \) for \( t \in [0, 5] \). Find the velocity when \( t = 3 \).

9. Let \( \mathbf{r}(t) = (\sin t, t^2, t + 1) \) for \( t \in [0, 5] \). Find the velocity when \( t = 3 \).

10. Let \( \mathbf{r}(t) = (t, \ln (t^2 + 1), t + 1) \) for \( t \in [0, 5] \). Find the velocity when \( t = 3 \).

11. Suppose an object has position \( \mathbf{r}(t) \in \mathbb{R}^3 \) where \( \mathbf{r} \) is differentiable and suppose also that \( |\mathbf{r}(t)| = c \) where \( c \) is a constant.

   (a) Show first that this condition does not require \( \mathbf{r}(t) \) to be a constant. **Hint:** You can do this either mathematically or by giving a physical example.

   (b) Show that you can conclude that \( \mathbf{r}'(t) \cdot \mathbf{r}(t) = 0 \). That is, the velocity is always perpendicular to the displacement.

12. Prove (13.5) from the component description of the cross product.

13. Prove (13.5) from the formula \((\mathbf{f} \times \mathbf{g})_i = \varepsilon_{ijk} f_j g_k\).

14. Prove (13.5) directly from the definition of the derivative without considering components.

15. A bezier curve in \( \mathbb{R}^n \) is a vector valued function of the form

\[
\mathbf{y}(t) = \sum_{k=0}^{n} \binom{n}{k} \mathbf{x}_k (1-t)^{n-k} t^k
\]

where here the \( \binom{n}{k} \) are the binomial coefficients and \( \mathbf{x}_k \) are \( n + 1 \) points in \( \mathbb{R}^n \). Show that \( \mathbf{y}(0) = \mathbf{x}_0 \), \( \mathbf{y}(1) = \mathbf{x}_n \), and find \( \mathbf{y}'(0) \) and \( \mathbf{y}'(1) \). Recall that \( \binom{n}{0} = \binom{n}{n} = 1 \) and \( \binom{n}{n-1} = \binom{n}{1} = n \). Curves of this sort are important in various computer programs.

16. Suppose \( \mathbf{r}(t), \mathbf{s}(t), \) and \( \mathbf{p}(t) \) are three differentiable functions of \( t \) which have values in \( \mathbb{R}^3 \). Find a formula for \( (\mathbf{r}(t) \times \mathbf{s}(t)) \cdot \mathbf{p}(t) \).

17. If \( \mathbf{r}'(t) = \mathbf{0} \) for all \( t \in (a, b) \), show there exists a constant vector, \( \mathbf{c} \) such that \( \mathbf{r}(t) = \mathbf{c} \) for all \( t \in (a, b) \).
13.6 Newton's Laws Of Motion

Definition 13.6.1 Let $r(t)$ denote the position of an object. Then the acceleration of the object is defined to be $r''(t)$.

Newton's first law is: “Every body persists in its state of rest or of uniform motion in a straight line unless it is compelled to change that state by forces impressed on it.”

Newton’s second law is:

$$F = ma$$

(13.9)

where $a$ is the acceleration and $m$ is the mass of the object.

Newton’s third law states: “To every action there is always opposed an equal reaction; or, the mutual actions of two bodies upon each other are always equal, and directed to contrary parts.”

Of these laws, only the second two are independent of each other, the first law being implied by the second. The third law says roughly that if you apply a force to something, the thing applies the same force back.

The second law is the one of most interest. Note that the statement of this law depends on the concept of the derivative because the acceleration is defined as a derivative. Newton used calculus and these laws to solve profound problems involving the motion of the planets and other problems in mechanics. The next example involves the concept that if you know the force along with the initial velocity and initial position, then you can determine the position.

Example 13.6.2 Let $r(t)$ denote the position of an object of mass 2 kilogram at time $t$ and suppose the force acting on the object is given by $F(t) = (t, 1-t^2, 2e^{-t})$. Suppose $r(0) = (1, 0, 1)$ meters, and $r'(0) = (0, 1, 1)$ meters/sec. Find $r(t)$.

By Newton’s second law, $2r''(t) = F(t) = (t, 1-t^2, 2e^{-t})$ and so

$$r''(t) = (t/2, (1-t^2)/2, e^{-t})$$

Therefore the velocity is given by

$$r'(t) = \left(\frac{t^2}{4}, \frac{t-t^3/3}{2}, -e^{-t}\right) + c$$
where \( c \) is a constant vector which must be determined from the initial condition given for the velocity. Thus letting \( c = (c_1, c_2, c_3) \),
\[
(0, 1, 1) = (0, 0, -1) + (c_1, c_2, c_3)
\]
which requires \( c_1 = 0, c_2 = 1, \) and \( c_3 = 2 \). Therefore, the velocity is found.
\[
r'(t) = \left( \frac{t^2}{4}, \frac{t - t^3/3}{2} + 1, -e^{-t} + 2 \right).
\]
Now from this, the displacement must equal
\[
r(t) = \left( \frac{t^3}{12}, \frac{t^2/2 - t^4/12}{2} + t, e^{-t} + 2t \right) + (C_1, C_2, C_3)
\]
where the constant vector, \((C_1, C_2, C_3)\) must be determined from the initial condition for the displacement. Thus
\[
r(0) = (1, 0, 1) = (0, 0, 1) + (C_1, C_2, C_3)
\]
which means \( C_1 = 1, C_2 = 0, \) and \( C_3 = 0 \). Therefore, the displacement has also been found.
\[
r(t) = \left( \frac{t^3}{12}, \frac{t^2/2 - t^4/12}{2} + t, e^{-t} + 2t \right) \text{ meters.}
\]
Actually, in applications of this sort of thing acceleration does not usually come to you as a nice given function written in terms of simple functions you understand. Rather, it comes as measurements taken by instruments and the position is continuously being updated based on this information. Another situation which often occurs is the case when the forces on the object depend not just on time but also on the position or velocity of the object.

**Example 13.6.3** An artillery piece is fired at ground level on a level plain. The angle of elevation is \( \pi/6 \) radians and the speed of the shell is 400 meters per second. How far does the shell fly before hitting the ground?

Neglect air resistance in this problem. Also let the direction of flight be along the positive \( x \) axis. Thus the initial velocity is the vector, \( 400 \cos (\pi/6) \mathbf{i} + 400 \sin (\pi/6) \mathbf{j} \) while the only force experienced by the shell after leaving the artillery piece is the force of gravity, \(-mg \mathbf{j}\) where \( m \) is the mass of the shell. The acceleration of gravity equals 9.8 meters per sec\(^2\) and so the following needs to be solved.

\[
mr''(t) = -mg \mathbf{j}, \quad r(0) = (0, 0), \quad r'(0) = 400 \cos (\pi/6) \mathbf{i} + 400 \sin (\pi/6) \mathbf{j}.
\]

Denoting \( r(t) \) as \((x(t), y(t))\),
\[
x''(t) = 0, \quad y''(t) = -g.
\]

Therefore, \( y'(t) = -gt + C \) and from the information on the initial velocity, \( C = 400 \sin (\pi/6) = 200 \). Thus
\[
y(t) = -4.9t^2 + 200t + D.
\]
\( D = 0 \) because the artillery piece is fired at ground level which requires both \( x \) and \( y \) to equal zero at this time. Similarly, \( x'(t) = 400 \cos (\pi/6) \) so \( x(t) = 400 \cos (\pi/6) t = 200 \sqrt{3} t \).

The shell hits the ground when \( y = 0 \) and this occurs when \(-4.9t^2 + 200t = 0 \). Thus \( t = 40/816.3265306 \) seconds and so at this time,
\[
x = 200 \sqrt{3}(40.816.3265306) = 14139.1902659 \text{ meters.}
\]

The next example is more complicated because it also takes in to account air resistance. We do not live in a vacuum.
Example 13.6.4 A lump of “blue ice” escapes the lavatory of a jet flying at 600 miles per hour at an altitude of 30,000 feet. This blue ice weighs 64 pounds near the earth and experiences a force of air resistance equal to \((-1.1)r(t)\) pounds. Find the position and velocity of the blue ice as a function of time measured in seconds. Also find the velocity when the lump hits the ground. Such lumps have been known to surprise people on the ground.

The first thing needed is to obtain information which involves consistent units. The blue ice weighs 32 pounds near the earth. Thus 32 pounds is the force exerted by gravity on the lump and so its mass must be given by Newton’s second law as follows.

\[
m = \frac{64}{32} = 2 \text{ slugs.}
\]

Thus \(m = 2\) slugs. The slug is the unit of mass in the system involving feet and pounds. The jet is flying at 600 miles per hour. I want to change this to feet per second. Thus it flies at

\[
\frac{600 \times 5280}{60 \times 60} = 880 \text{ feet per second.}
\]

The explanation for this is that there are 5280 feet in a mile and so it goes \(600 \times 5280\) feet in one hour. There are \(60 \times 60\) seconds in an hour. The position of the lump of blue ice will be computed from a point on the ground directly beneath the airplane at the instant the blue ice escapes and regard the airplane as moving in the direction of the positive \(x\) axis. Thus the initial displacement is

\[
r(0) = (0, 30000) \text{ feet}
\]

and the initial velocity is

\[
r'(0) = (880, 0) \text{ feet/sec.}
\]

The force of gravity is

\[
(0, -64) \text{ pounds}
\]

and the force due to air resistance is

\[
(-1.1)r'(t) \text{ pounds.}
\]

Newton’s second law yields the following initial value problem for \(r(t) = (r_1(t), r_2(t))\).

\[
2(r''_1(t), r''_2(t)) = (-1.1)(r'_1(t), r'_2(t)) + (0, -64), (r_1(0), r_2(0)) = (0, 30000), \quad (r'_1(0), r'_2(0)) = (880, 0)
\]

Therefore,

\[
\begin{align*}
2r''_1(t) + (1.1)r'_1(t) &= 0 \\
2r''_2(t) + (1.1)r'_2(t) &= -64 \\
r_1(0) &= 0 \\
r_2(0) &= 30000 \\
r'_1(0) &= 880 \\
r'_2(0) &= 0
\end{align*}
\]

(13.10)

To save on repetition solve

\[
mr'' + kr' = c, r(0) = u, r'(0) = v.
\]

Divide the differential equation by \(m\) and get

\[
r'' + \left(\frac{k}{m}\right)r' = \frac{c}{m}.
\]
Now multiply both sides by \( e^{(k/m)t} \). You should check this gives
\[
\frac{d}{dt} \left( e^{(k/m)t} r' \right) = (c/m) e^{(k/m)t}
\]
Therefore,
\[
e^{(k/m)t} r' = \frac{1}{k} e^{\frac{k}{m}} t + C
\]
and using the initial condition, \( v = c/k + C \) so
\[
r'(t) = (c/k) + (v - (c/k)) e^{-\frac{k}{m} t}
\]
Now this implies
\[
r(t) = (c/k) t - \frac{1}{k} m e^{-\frac{k}{m} t} \left( v - \frac{c}{k} \right) + D
\]
where \( D \) is a constant to be determined from the initial conditions. Thus
\[
u = -\frac{m}{k} \left( v - \frac{c}{k} \right) + D
\]
and so
\[
r(t) = (c/k) t - \frac{1}{k} m e^{-\frac{k}{m} t} \left( v - \frac{c}{k} \right) + \left( u + \frac{m}{k} \left( v - \frac{c}{k} \right) \right).
\]
Now apply this to the system \((13.10)\) to find
\[
r_1(t) = -\frac{1}{(.1)^2} 2 \left( \exp \left( \frac{-(.1)^2}{2} \right) \right) (880) + \left( \frac{2}{(.1)} (880) \right)
\]
\[= -17600.0 \exp (-.05t) + 17600.0
\]
and
\[
r_2(t) = (-64/(.1)) t - \frac{1}{(.1)^2} \left( \exp \left( \frac{-(.1)^2}{2} \right) \right) \left( \frac{64}{(.1)} \right) + \left( 30000 + \frac{2}{(.1)} \left( \frac{64}{(.1)} \right) \right)
\]
\[= -640.0 t - 12800.0 \exp (-.05t) + 42800.0
\]
This gives the coordinates of the position. What of the velocity? Using \((13.11)\) in the same way to obtain the velocity,
\[
r_1'(t) = 880.0 \exp (-.05t),
nr_2'(t) = -640.0 + 640.0 \exp (-.05t).
\]
To determine the velocity when the blue ice hits the ground, it is necessary to find the value of \( t \) when this event takes place and then to use \((13.12)\) to determine the velocity. It hits ground when \( r_2(t) = 0 \). Thus it suffices to solve the equation,
\[0 = -640.0 t - 12800.0 \exp (-.05t) + 42800.0.
\]
This is a fairly hard equation to solve using the methods of algebra. In fact, I do not have a good way to find this value of \( t \) using algebra. However if plugging in various values of \( t \) using a calculator you eventually find that when \( t = 66.14 \),
\[-640.0 (66.14) - 12800.0 \exp (-.05 (66.14)) + 42800.0 = 1.588 \text{ feet}.
\]
This is close enough to hitting the ground and so plugging in this value for \( t \) yields the approximate velocity,
\[(880.0 \exp (-.05 (66.14)), -640.0 + 640.0 \exp (-.05 (66.14))) = (32.23, -616.56).\]
Notice how because of air resistance the component of velocity in the horizontal direction is only about 32 feet per second even though this component started out at 880 feet per second while the component in the vertical direction is -616 feet per second even though this component started off at 0 feet per second. You see that air resistance can be very important so it is not enough to pretend, as is often done in beginning physics courses that everything takes place in a vacuum. Actually, this problem used several physical simplifications. It was assumed the force acting on the lump of blue ice by gravity was constant. This is not really true because it actually depends on the distance between the center of mass of the earth and the center of mass of the lump. It was also assumed the air resistance is proportional to the velocity. This is an over simplification when high speeds are involved. However, increasingly correct models can be studied in a systematic way as above.

### 13.6.1 Kinetic Energy

Newton’s second law is also the basis for the notion of kinetic energy. When a force is exerted on an object which causes the object to move, it follows that the force is doing work which manifests itself in a change of velocity of the object. How is the total work done on the object by the force related to the final velocity of the object? By Newton’s second law, and letting $v$ be the velocity,

$$F(t) = m v'(t).$$

Now in a small increment of time, $(t, t + dt)$, the work done on the object would be approximately equal to

$$dW = F(t) \cdot v(t) \, dt. \quad (13.13)$$

If no work has been done at time $t = 0$, then (13.13) implies

$$\frac{dW}{dt} = F(t) \cdot v(t), \quad W(0) = 0.$$

Hence,

$$\frac{dW}{dt} = m v'(t) \cdot v(t) = \frac{m}{2} \frac{d}{dt} |v(t)|^2.$$

Therefore, the total work done up to time $t$ would be $W(t) = \frac{m}{2} |v(t)|^2 - \frac{m}{2} |v_0|^2$ where $|v_0|$ denotes the initial speed of the object. This difference represents the change in the kinetic energy.

### 13.6.2 Impulse And Momentum

Work and energy involve a force acting on an object for some distance. Impulse involves a force which acts on an object for an interval of time.

**Definition 13.6.5** Let $F$ be a force which acts on an object during the time interval, $[a, b]$. The impulse of this force is

$$\int_a^b F(t) \, dt.$$

This is defined as

$$\left( \int_a^b F_1(t) \, dt, \int_a^b F_2(t) \, dt, \int_a^b F_3(t) \, dt \right).$$

The linear momentum of an object of mass $m$ and velocity $v$ is defined as

$$\text{Linear momentum} = mv.$$
The notion of impulse and momentum are related in the following theorem.

**Theorem 13.6.6** Let \( \mathbf{F} \) be a force acting on an object of mass \( m \). Then the impulse equals the change in momentum. More precisely,

\[
\int_{a}^{b} \mathbf{F}(t) \, dt = m \mathbf{v}(b) - m \mathbf{v}(a).
\]

**Proof:** This is really just the fundamental theorem of calculus and Newton’s second law applied to the components of \( \mathbf{F} \).

\[
\int_{a}^{b} \mathbf{F}(t) \, dt = \int_{a}^{b} m \frac{d\mathbf{v}}{dt} \, dt = m \mathbf{v}(b) - m \mathbf{v}(a) \quad (13.14)
\]

Now suppose two point masses, \( A \) and \( B \) collide. Newton’s third law says the force exerted by mass \( A \) on mass \( B \) is equal in magnitude but opposite in direction to the force exerted by mass \( B \) on mass \( A \). Letting the collision take place in the time interval, \([a, b]\) and denoting the two masses by \( m_A \) and \( m_B \) and their velocities by \( \mathbf{v}_A \) and \( \mathbf{v}_B \) it follows that

\[
m_A \mathbf{v}_A(b) - m_A \mathbf{v}_A(a) = \int_{a}^{b} \text{(Force of } B \text{ on } A) \, dt
\]

and

\[
m_B \mathbf{v}_B(b) - m_B \mathbf{v}_B(a) = \int_{a}^{b} \text{(Force of } A \text{ on } B) \, dt
\]

\[
= - \int_{a}^{b} \text{(Force of } B \text{ on } A) \, dt
\]

\[
= - (m_A \mathbf{v}_A(b) - m_A \mathbf{v}_A(a))
\]

and this shows

\[
m_B \mathbf{v}_B(b) + m_A \mathbf{v}_A(b) = m_B \mathbf{v}_B(a) + m_A \mathbf{v}_A(a).
\]

In other words, in a collision between two masses the total linear momentum before the collision equals the total linear momentum after the collision. This is known as the conservation of linear momentum.

### 13.7 Moving Coordinate Systems

The idea is you have a coordinate system which is moving and this results in strange forces experienced relative to these moving coordinate systems. A good example is what we experience every day living on a rotating ball. Relative to our supposedly fixed coordinate system, we experience forces which account for many phenomena which are observed.

#### 13.7.1 The Coriolis Acceleration

Imagine a point on the surface of the earth. Now consider unit vectors, one pointing South, one pointing East and one pointing directly away from the center of the earth.
Denote the first as \( i \), the second as \( j \) and the third as \( k \). If you are standing on the earth you will consider these vectors as fixed, but of course they are not. As the earth turns, they change direction and so each is in reality a function of \( t \). Nevertheless, it is with respect to these apparently fixed vectors that you wish to understand acceleration, velocities, and displacements.

In general, let \( \mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t) \) be an orthonormal basis of vectors for each \( t \), like the vectors described in the first paragraph. It is assumed these vectors are \( C^1 \) functions of \( t \). Letting the positive \( x \) axis extend in the direction of \( \mathbf{i}(t) \), the positive \( y \) axis extend in the direction of \( \mathbf{j}(t) \), and the positive \( z \) axis extend in the direction of \( \mathbf{k}(t) \), yields a moving coordinate system. By Theorem 13.4.2 on Page 241, there exists an angular velocity vector, \( \mathbf{\Omega}(t) \) such that if \( \mathbf{u}(t) \) is any vector which has constant components with respect to \( \mathbf{i}(t), \mathbf{j}(t), \) and \( \mathbf{k}(t) \), then

\[
\mathbf{\Omega} \times \mathbf{u} = \mathbf{u}'.
\]  

(13.15)

Now let \( \mathbf{R}(t) \) be a position vector of the moving coordinate system and let

\[
\mathbf{r}(t) = \mathbf{R}(t) + \mathbf{r}_B(t)
\]

where

\[
\mathbf{r}_B(t) \equiv x(t) \mathbf{i}(t) + y(t) \mathbf{j}(t) + z(t) \mathbf{k}(t).
\]

In the example of the earth, \( \mathbf{R}(t) \) is the position vector of a point \( p(t) \) on the earth’s surface and \( \mathbf{r}_B(t) \) is the position vector of another point from \( p(t) \), thus regarding \( \mathbf{p}(t) \) as the origin. \( \mathbf{r}_B(t) \) is the position vector of a point as perceived by the observer on the earth with respect to the vectors he thinks of as fixed. Similarly, \( \mathbf{v}_B(t) \) and \( \mathbf{a}_B(t) \) will be the velocity and acceleration relative to \( \mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t) \), and so \( \mathbf{v}_B = x'i + y'j + z'k \) and \( \mathbf{a}_B = x''i + y''j + z''k \). Then

\[
\mathbf{v} \equiv \mathbf{v}' = \mathbf{v}' + x'i + y'j + z'k + x'i' + y'j' + z'k'.
\]

By (13.15), if \( \mathbf{e} \in \{ \mathbf{i}, \mathbf{j}, \mathbf{k} \} \), \( \mathbf{e}' = \mathbf{\Omega} \times \mathbf{e} \) because the components of these vectors with respect to \( \mathbf{i}, \mathbf{j}, \mathbf{k} \) are constant. Therefore,

\[
x'i' + y'j' + z'k' = x\mathbf{\Omega} \times \mathbf{i} + y\mathbf{\Omega} \times \mathbf{j} + z\mathbf{\Omega} \times \mathbf{k}
\]

\[
= \mathbf{\Omega} \times (xi + yj + zk)
\]
and consequently,

\[ \mathbf{v} = \mathbf{R}' + x'i + y'j + z'k + \mathbf{\Omega} \times \mathbf{r}_B = \mathbf{R}' + x'i + y'j + z'k + \mathbf{\Omega} \times (xi + yj + zk). \]

Now consider the acceleration. Quantities which are relative to the moving coordinate system are distinguished by using the subscript, \( B \).

\[
\mathbf{a} = \mathbf{v}' = \mathbf{R}'' + x''i + y''j + z''k + \mathbf{\Omega} \times \mathbf{v}_B + \mathbf{\Omega} \times \mathbf{r}_B
\]

\[
+ \mathbf{\Omega} \times \left( x'i + y'j + z'k + xi + yj + zk \right)
\]

\[ = \mathbf{R}'' + \mathbf{a}_B + \mathbf{\Omega} \times \mathbf{r}_B + 2\mathbf{\Omega} \times \mathbf{v}_B + \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}_B). \]

The acceleration \( \mathbf{a}_B \) is that perceived by an observer for whom the moving coordinate system is fixed. The term \( \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}_B) \) is called the centripetal acceleration. Solving for \( \mathbf{a}_B \),

\[
\mathbf{a}_B = \mathbf{a} - \mathbf{R}'' - \mathbf{\Omega} \times \mathbf{r}_B - 2\mathbf{\Omega} \times \mathbf{v}_B - \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}_B). \quad (13.16)
\]

Here the term \( - (\mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}_B)) \) is called the centrifugal acceleration, it being an acceleration felt by the observer relative to the moving coordinate system which he regards as fixed, and the term \( - 2\mathbf{\Omega} \times \mathbf{v}_B \) is called the Coriolis acceleration, an acceleration experienced by the observer as he moves relative to the moving coordinate system. The mass multiplied by the Coriolis acceleration defines the Coriolis force.

There is a ride found in some amusement parks in which the victims stand next to a circular wall covered with a carpet or some rough material. Then the whole circular room begins to revolve faster and faster. At some point, the bottom drops out and the victims are held in place by friction. The force they feel which keeps them stuck to the wall is called centrifugal force and it causes centrifugal acceleration. It is not necessary to move relative to coordinates fixed with the revolving wall in order to feel this force and it is pretty predictable. However, if the nauseated victim moves relative to the rotating wall, he will feel the effects of the Coriolis force and this force is really strange. The difference between these forces is that the Coriolis force is caused by movement relative to the moving coordinate system and the centrifugal force is not.

### 13.7.2 The Coriolis Acceleration On The Rotating Earth

Now consider the earth. Let \( \mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^* \), be the usual basis vectors attached to the rotating earth. Thus \( \mathbf{k}^* \) is fixed in space with \( \mathbf{k}^* \) pointing in the direction of the north pole from the center of the earth while \( \mathbf{i}^* \) and \( \mathbf{j}^* \) point to fixed points on the surface of the earth. Thus \( \mathbf{i}^* \) and \( \mathbf{j}^* \) depend on \( t \) while \( \mathbf{k}^* \) does not. Let \( \mathbf{i}, \mathbf{j}, \mathbf{k} \) be the unit vectors described earlier with \( \mathbf{i} \) pointing South, \( \mathbf{j} \) pointing East, and \( \mathbf{k} \) pointing away from the center of the earth at some point of the rotating earth’s surface, \( \mathbf{p} \). Letting \( \mathbf{R}(t) \) be the position vector of the point \( \mathbf{p} \), from the center of the earth, observe the coordinates of \( \mathbf{R}(t) \) are constant with respect to \( \mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t) \). Also, since the earth rotates from West to East and the speed of a point on the surface of the earth relative to an observer fixed in space is \( \omega |\mathbf{R}| \sin \phi \) where \( \omega \) is the angular speed of the earth about an axis through the poles, it follows from the geometric definition of the cross product that

\[
\mathbf{R}' = \omega \mathbf{k}^* \times \mathbf{R}
\]
Therefore, $\Omega = \omega k^*$ and so

$$R'' = \Omega' \times \dot{R} + \Omega \times R' = \Omega \times (\Omega \times R)$$

since $\Omega$ does not depend on $t$. Formula (13.16) implies

$$a_B = a - \Omega \times (\Omega \times R) - 2\Omega \times v_B - \Omega \times (\Omega \times r_B).$$

(13.17)

In this formula, you can totally ignore the term $\Omega \times (\Omega \times r_B)$ because it is so small whenever you are considering motion near some point on the earth’s surface. To see this, note

$$\omega (24)(3600) = 2\pi,$$

and so $\omega = 7.2722 \times 10^{-5}$ in radians per second. If you are using seconds to measure time and feet to measure distance, this term is therefore, no larger than

$$(7.2722 \times 10^{-5})^2 |r_B|.$$ 

Clearly this is not worth considering in the presence of the acceleration due to gravity which is approximately 32 feet per second squared near the surface of the earth.

If the acceleration $a$, is due to gravity, then

$$a_B = a - \Omega \times (\Omega \times R) - 2\Omega \times v_B =$$

$$= g - \frac{GM (R + r_B)}{|R + r_B|^3} - \Omega \times (\Omega \times R) - 2\Omega \times v_B \equiv g - 2\Omega \times v_B.$$ 

Note that

$$\Omega \times (\Omega \times R) = (\Omega \cdot R) \Omega - |\Omega|^2 R$$

and so $g$, the acceleration relative to the moving coordinate system on the earth is not directed exactly toward the center of the earth except at the poles and at the equator, although the components of acceleration which are in other directions are very small when compared with the acceleration due to the force of gravity and are often neglected. Therefore, if the only force acting on an object is due to gravity, the following formula describes the acceleration relative to a coordinate system moving with the earth’s surface.

$$a_B = g - 2(\Omega \times v_B)$$

While the vector, $\Omega$ is quite small, if the relative velocity, $v_B$ is large, the Coriolis acceleration could be significant. This is described in terms of the vectors $i(t), j(t), k(t)$ next.

Letting $(\rho, \theta, \phi)$ be the usual spherical coordinates of the point $p(t)$ on the surface taken with respect to $i^*, j^*, k^*$ the usual way with $\phi$ the polar angle, it follows the $i^*, j^*, k^*$ coordinates of this point are

$$\begin{pmatrix}
\rho 
\sin (\phi) \cos (\theta) \\
\rho 
\sin (\phi) \sin (\theta) \\
\rho 
\cos (\phi)
\end{pmatrix}.$$ 

It follows,

$$i = \cos (\phi) \cos (\theta) i^* + \cos (\phi) \sin (\theta) j^* - \sin (\phi) k^*$$

$$j = - \sin (\theta) i^* + \cos (\theta) j^* + 0 k^*$$

and

$$k = \sin (\phi) \cos (\theta) i^* + \sin (\phi) \sin (\theta) j^* + \cos (\phi) k^*.$$
It is necessary to obtain $k^*$ in terms of the vectors, $i, j, k$. Thus the following equation needs to be solved for $a, b, c$ to find $k^* = a_i + b_j + c_k$

\[
\begin{pmatrix}
0 \\
0 \\
1 \\
\end{pmatrix} = \begin{pmatrix}
\cos(\phi) \cos(\theta) & -\sin(\theta) & \sin(\phi) \cos(\theta) \\
\cos(\phi) \sin(\theta) & \cos(\phi) \sin(\theta) & 0 \\
-\sin(\phi) & 0 & \cos(\phi) \\
\end{pmatrix} \begin{pmatrix}
a \\
b \\
c \\
\end{pmatrix}
\] (13.18)

The first column is $i$, the second is $j$ and the third is $k$ in the above matrix. The solution is $a = -\sin(\phi), b = 0,$ and $c = \cos(\phi)$.

Now the Coriolis acceleration on the earth equals

\[
2 (\Omega \times v_B) = 2\omega \begin{pmatrix}
0 \\
0 \\
1 \\
\end{pmatrix} \times (x'i + y'j + z'k).
\]

This equals

\[
2\omega \left[ (-y' \cos \phi) i + (x' \cos \phi + z' \sin \phi) j - (y' \sin \phi) k \right].
\] (13.19)

Remember $\phi$ is fixed and pertains to the fixed point, $p(t)$ on the earth’s surface. Therefore, if the acceleration, $a$, is due to gravity,

\[
a_B = g - 2\omega \left[ (-y' \cos \phi) i + (x' \cos \phi + z' \sin \phi) j - (y' \sin \phi) k \right]
\]

where $g = \frac{-GM(R+r_B)}{(R+r_B)^3} - \Omega \times (\Omega \times R)$ as explained above. The term $\Omega \times (\Omega \times R)$ is pretty small and so it will be neglected. However, the Coriolis force will not be neglected.

**Example 13.7.1** Suppose a rock is dropped from a tall building. Where will it strike?

Assume $a = -gk$ and the $j$ component of $a_B$ is approximately

\[-2\omega (x' \cos \phi + z' \sin \phi).
\]

The dominant term in this expression is clearly the second one because $x'$ will be small. Also, the $i$ and $k$ contributions will be very small. Therefore, the following equation is descriptive of the situation.

\[a_B = -gk - 2z' \omega \sin \phi j.
\]

$z' = -gt$ approximately. Therefore, considering the $j$ component, this is

\[2gt \omega \sin \phi.
\]

Two integrations give $(\omega gt^3/3) \sin \phi$ for the $j$ component of the relative displacement at time $t$.

This shows the rock does not fall directly towards the center of the earth as expected but slightly to the east.

**Example 13.7.2** In 1851 Foucault set a pendulum vibrating and observed the earth rotate out from under it. It was a very long pendulum with a heavy weight at the end so that it would vibrate for a long time without stopping. This is what allowed him to observe the earth rotate out from under it. Clearly such a pendulum will take 24 hours for the plane of vibration to appear to make one complete revolution at the north pole. It is also reasonable to expect that no such observed rotation would take place on the equator. Is it possible to predict what will take place at various latitudes?

\[^{3}\text{There is such a pendulum in the Eyring building at BYU and to keep people from touching it, there is a little sign which says Warning! 1000 ohms.}\]
Using (13.19), in (13.17),
\[ a_B = a - \Omega \times (\Omega \times R) \]
\[-2\omega \left[ (-y' \cos \phi) \mathbf{i} + (x' \cos \phi + z' \sin \phi) \mathbf{j} - (y' \sin \phi) \mathbf{k} \right].\]
Neglecting the small term, \( \Omega \times (\Omega \times R) \), this becomes
\[ = -g \mathbf{k} + \frac{T}{m} - 2\omega \left[ (-y' \cos \phi) \mathbf{i} + (x' \cos \phi + z' \sin \phi) \mathbf{j} - (y' \sin \phi) \mathbf{k} \right] \]
where \( T \), the tension in the string of the pendulum, is directed towards the point at which the pendulum is supported, and \( m \) is the mass of the pendulum bob. The pendulum can be thought of as the position vector from \((0,0,l)\) to the surface of the sphere \( x^2 + y^2 + (z - l)^2 = l^2 \). Therefore,
\[ T = -T \frac{x}{l} \mathbf{i} - T \frac{y}{l} \mathbf{j} + T \frac{l - z}{l} \mathbf{k} \]
and consequently, the differential equations of relative motion are
\[ x'' = -\frac{g}{l} x + 2\omega y' \cos \phi \]
\[ y'' = -\frac{g}{l} y - 2\omega (x' \cos \phi + z' \sin \phi) \]
and
\[ z'' = \frac{g}{l} - z + 2\omega y' \sin \phi. \]
If the vibrations of the pendulum are small so that for practical purposes, \( z'' = z = 0 \), the last equation may be solved for \( T \) to get
\[ gm - 2\omega y' \sin (\phi) m = T. \]
Therefore, the first two equations become
\[ x'' = - (gm - 2\omega my' \sin \phi) \frac{x}{ml} + 2\omega y' \cos \phi \]
and
\[ y'' = - (gm - 2\omega my' \sin \phi) \frac{y}{ml} - 2\omega (x' \cos \phi + z' \sin \phi). \]
All terms of the form \( xy' \) or \( y'y \) can be neglected because it is assumed \( x \) and \( y \) remain small. Also, the pendulum is assumed to be long with a heavy weight so that \( x' \) and \( y' \) are also small. With these simplifying assumptions, the equations of motion become
\[ x'' + g \frac{x}{l} = 2\omega y' \cos \phi \]
and
\[ y'' + g \frac{y}{l} = -2\omega x' \cos \phi. \]
These equations are of the form
\[ x'' + a^2 x = by', \ y'' + a^2 y = -bx' \]
where \( a^2 = \frac{g}{l} \) and \( b = 2\omega \cos \phi \). Then it is fairly tedious but routine to verify that for each constant, \( c \),
\[ x = c \sin \left( \frac{bt}{2} \right) \sin \left( \frac{\sqrt{b^2 + 4a^2} t}{2} \right), \ y = c \cos \left( \frac{bt}{2} \right) \sin \left( \frac{\sqrt{b^2 + 4a^2} t}{2} \right) \]
yields a solution to (13.20) along with the initial conditions,
\[ x(0) = 0, y(0) = 0, x'(0) = 0, y'(0) = \frac{c\sqrt{b^2 + 4a^2}}{2}. \]

(13.22)

It is clear from experiments with the pendulum that the earth does indeed rotate out from under it causing the plane of vibration of the pendulum to appear to rotate. The purpose of this discussion is not to establish these self evident facts but to predict how long it takes for the plane of vibration to make one revolution. Therefore, there will be some instant in time at which the pendulum will be vibrating in a plane determined by \( k \) and \( j \). (Recall \( k \) points away from the center of the earth and \( j \) points East.) At this instant in time, defined as \( t = 0 \), the conditions of (13.22) will hold for some value of \( c \) and so the solution to (13.20) having these initial conditions will be those of (13.21) by uniqueness of the initial value problem. Writing these solutions differently,
\[
\begin{pmatrix}
  x(t) \\
  y(t)
\end{pmatrix} = c \left( \sin \left( \frac{bt}{2} \right) \right) \sin \left( \frac{\sqrt{b^2 + 4a^2}}{2} t \right)
\]

This is very interesting! The vector, \( c \left( \sin \left( \frac{bt}{2} \right) \right) \) always has magnitude equal to \( |c| \) but its direction changes very slowly because \( b \) is very small. The plane of vibration is determined by this vector and the vector \( k \). The term \( \sin \left( \frac{\sqrt{b^2 + 4a^2}}{2} t \right) \) changes relatively fast and takes values between \(-1\) and \(1\). This is what describes the actual observed vibrations of the pendulum. Thus the plane of vibration will have made one complete revolution when \( t = P \) for
\[
\frac{bP}{2} \equiv 2\pi.
\]

Therefore, the time it takes for the earth to turn out from under the pendulum is
\[
P = \frac{4\pi}{2\omega \cos \phi} = \frac{2\pi}{\omega} \sec \phi.
\]

Since \( \omega \) is the angular speed of the rotating earth, it follows \( \omega = \frac{2\pi}{24} = \frac{\pi}{12} \) in radians per hour. Therefore, the above formula implies
\[
P = 24 \sec \phi.
\]

I think this is really amazing. You could actually determine latitude, not by taking readings with instruments using the North Star but by doing an experiment with a big pendulum. You would set it vibrating, observe \( P \) in hours, and then solve the above equation for \( \phi \). Also note the pendulum would not appear to change its plane of vibration at the equator because \( \lim_{\phi \to \pi/2} \sec \phi = \infty \).

The Coriolis acceleration is also responsible for the phenomenon of the next example.

Example 13.7.3 It is known that low pressure areas rotate counterclockwise as seen from above in the Northern hemisphere but clockwise in the Southern hemisphere. Why?

Neglect accelerations other than the Coriolis acceleration and the following acceleration which comes from an assumption that the point \( p(t) \) is the location of the lowest pressure.
\[
a = -a(r_B) r_B
\]

where \( r_B \) will denote the distance from the fixed point \( p(t) \) on the earth’s surface which is also the lowest pressure point. Of course the situation could be more complicated but
this will suffice to explain the above question. Then the acceleration observed by a person on the earth relative to the apparently fixed vectors, $\mathbf{i}, \mathbf{k}, \mathbf{j}$, is

$$\mathbf{a}_B = -a(r_B)(x\mathbf{i} + y\mathbf{j} + z\mathbf{k}) - 2\omega [ -y' \cos(\phi) \mathbf{i} + (x' \cos(\phi) + z' \sin(\phi)) \mathbf{j} - (y' \sin(\phi)) \mathbf{k}]$$

Therefore, one obtains some differential equations from $\mathbf{a}_B = x''\mathbf{i} + y''\mathbf{j} + z''\mathbf{k}$ by matching the components. These are

$$x'' + a(r_B) x = 2\omega y' \cos \phi$$
$$y'' + a(r_B) y = -2\omega x' \cos \phi - 2\omega z' \sin \phi$$
$$z'' + a(r_B) z = 2\omega y' \sin \phi$$

Now remember, the vectors, $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are fixed relative to the earth and so are constant vectors. Therefore, from the properties of the determinant and the above differential equations,

$$\left(\mathbf{r}_B \times \mathbf{r}_B\right)' = \left| \begin{array}{ccc}
\mathbf{i} & \mathbf{j} & \mathbf{k} \\
x' & y' & z' \\
x & y & z 
\end{array} \right| = \left| \begin{array}{ccc}
\mathbf{i} & \mathbf{j} & \mathbf{k} \\
x'' & y'' & z'' \\
x & y & z 
\end{array} \right|$$

$$= -a(r_B)x + 2\omega y' \cos \phi - a(r_B)y - 2\omega x' \cos \phi - 2\omega z' \sin \phi - a(r_B)z + 2\omega y' \sin \phi$$

Then the $k^{th}$ component of this cross product equals

$$\omega \cos(\phi) \left( y'^2 + x'^2 \right)' + 2\omega x z' \sin(\phi).$$

The first term will be negative because it is assumed $\mathbf{p}(t)$ is the location of low pressure causing $y'^2 + x'^2$ to be a decreasing function. If it is assumed there is not a substantial motion in the $k$ direction, so that $z$ is fairly constant and the last term can be neglected, then the $k^{th}$ component of $(\mathbf{r}_B \times \mathbf{r}_B)'$ is negative provided $\phi \in (0, \pi/2)$ and positive if $\phi \in (\pi/2, \pi)$. Beginning with a point at rest, this implies $\mathbf{r}_B \times \mathbf{r}_B = 0$ initially and then the above implies its $k^{th}$ component is negative in the upper hemisphere when $\phi < \pi/2$ and positive in the lower hemisphere when $\phi > \pi/2$. Using the right hand and the geometric definition of the cross product, this shows clockwise rotation in the lower hemisphere and counter clockwise rotation in the upper hemisphere.

Note also that as $\phi$ gets close to $\pi/2$ near the equator, the above reasoning tends to break down because $\cos(\phi)$ becomes close to zero. Therefore, the motion towards the low pressure has to be more pronounced in comparison with the motion in the $k$ direction in order to draw this conclusion.

### 13.8 Exercises

1. Show the solution to $\mathbf{v}' + r \mathbf{v} = \mathbf{c}$ with the initial condition, $\mathbf{v}(0) = \mathbf{v}_0$ is $\mathbf{v}(t) = (v_0 - \frac{c}{r}) e^{-rt} + \left(\mathbf{c}/r\right)$. If $\mathbf{v}$ is velocity and $r = k/m$ where $k$ is a constant for air resistance and $m$ is the mass, and $\mathbf{c} = \mathbf{f}/m$, argue from Newton’s second law that this is the equation for finding the velocity, $\mathbf{v}$ of an object acted on by air resistance proportional to the velocity and a constant force, $\mathbf{f}$, possibly from gravity. Does there exist a terminal velocity? What is it?

2. Verify Formula (13.14) carefully by considering the components.

3. Suppose that the air resistance is proportional to the velocity but it is desired to find the constant of proportionality. Describe how you could find this constant.
4. Suppose an object having mass equal to 5 kilograms experiences a time dependent force, \( \mathbf{F}(t) = e^{-t}\mathbf{i} + \cos(t)\mathbf{j} + t^2\mathbf{k} \) meters per sec\(^2\). Suppose also that the object is at the point \((0, 1, 1)\) meters at time \(t = 0\) and that its initial velocity at this time is \(\mathbf{v} = \mathbf{i} + \mathbf{j} - \mathbf{k}\) meters per sec. Find the position of the object as a function of \(t\).

5. Fill in the details for the derivation of kinetic energy. In particular verify that \(m\mathbf{v}'(t) \cdot \mathbf{v}(t) = \frac{m}{2} \frac{d}{dt} |\mathbf{v}(t)|^2\). Also, why would \(dW = \mathbf{F}(t) \cdot \mathbf{v}(t) \, dt\)?

6. Suppose the force acting on an object, \(\mathbf{F}\) is always perpendicular to the velocity of the object. Thus \(\mathbf{F} \cdot \mathbf{v} = 0\). Show that the Kinetic energy of the object is constant. Such forces are sometimes called forces of constraint because they do not contribute to the speed of the object, only its direction.

7. A cannon is fired at an angle, \(\theta\) from ground level on a vast plain. The speed of the ball as it leaves the mouth of the cannon is known to be \(s\) meters per second. Neglecting air resistance, find a formula for how far the cannon ball goes before hitting the ground. Show the maximum range for the cannon ball is achieved when \(\theta = \pi/4\).

8. Suppose in the context of Problem 7 that the cannon ball has mass 10 kilograms and it experiences a force of air resistance which is \(0.01\mathbf{v}\) Newtons where \(\mathbf{v}\) is the velocity in meters per second. The acceleration of gravity is 9.8 meters per sec\(^2\). Also suppose that the initial speed is 100 meters per second. Find a formula for the displacement, \(\mathbf{r}(t)\) of the cannon ball. If the angle of elevation equals \(\pi/4\), use a calculator or other means to estimate the time before the cannon ball hits the ground.

9. Show that Newton’s first law can be obtained from the second law.

10. Show that if \(\mathbf{v}'(t) = \mathbf{0}\), for all \(t \in (a, b)\), then there exists a constant vector, \(\mathbf{z}\) independent of \(t\) such that \(\mathbf{v}(t) = \mathbf{z}\) for all \(t\).

11. Suppose an object moves in three dimensional space in such a way that the only force acting on the object is directed toward a single fixed point in three dimensional space. Verify that the motion of the object takes place in a plane. **Hint:** Let \(\mathbf{r}(t)\) denote the position vector of the object from the fixed point. Then the force acting on the object must be of the form \(g(\mathbf{r}(t))\mathbf{r}(t)\) and by Newton’s second law, this equals \(m\mathbf{r}''(t)\). Therefore,
\[
m\mathbf{r}'' \times \mathbf{r} = g(\mathbf{r})\mathbf{r} \times \mathbf{r} = \mathbf{0}.
\]
Now argue that \(\mathbf{r}'' \times \mathbf{r} = (\mathbf{r}' \times \mathbf{r})'\), showing that \((\mathbf{r}' \times \mathbf{r})\) must equal a constant vector, \(\mathbf{z}\). Therefore, what can be said about \(\mathbf{z}\) and \(\mathbf{r}\)?

12. Suppose the only forces acting on an object are the force of gravity, \(-mg\mathbf{k}\) and a force, \(\mathbf{F}\) which is perpendicular to the motion of the object. Thus \(\mathbf{F} \cdot \mathbf{v} = 0\). Show the total energy of the object,
\[
E \equiv \frac{1}{2}m|\mathbf{v}|^2 + mgz
\]
is constant. Here \(\mathbf{v}\) is the velocity and the first term is the kinetic energy while the second is the potential energy. **Hint:** Use Newton’s second law to show the time derivative of the above expression equals zero.

13. Using Problem 12, suppose an object slides down a frictionless inclined plane from a height of 100 feet. When it reaches the bottom, how fast will it be going? Assume it starts from rest.
14. The ballistic pendulum is an interesting device which is used to determine the speed of a bullet. It is a large massive block of wood hanging from a long string. A rifle is fired into the block of wood which then moves. The speed of the bullet can be determined from measuring how high the block of wood rises. Explain how this can be done and why. **Hint:** Let \( v \) be the speed of the bullet which has mass \( m \) and let the block of wood have mass \( M \). By conservation of momentum \( mv = (m + M)V \) where \( V \) is the speed of the block of wood immediately after the collision. Thus the energy is \( \frac{1}{2} (m + M)V^2 \) and this block of wood rises to a height of \( h \). Now use Problem 12.

15. In the experiment of Problem 14, show the kinetic energy before the collision is greater than the kinetic energy after the collision. Thus linear momentum is conserved but energy is not. Such a collision is called inelastic.

16. There is a popular toy consisting of identical steel balls suspended from strings of equal length as illustrated in the following picture.

![Toy Illustration](image)

The ball at the right is lifted and allowed to swing. When it collides with the other balls, the ball on the left is observed to swing away from the others with the same speed the ball on the right had when it collided. Why does this happen? Why don’t two or more of the stationary balls start to move, perhaps at a slower speed? This is an example of an elastic collision because energy is conserved. Of course this could change if you fixed things so the balls would stick to each other.

17. An illustration used in many beginning physics books is that of firing a rifle horizontally and dropping an identical bullet from the same height above the perfectly flat ground followed by an assertion that the two bullets will hit the ground at exactly the same time. Is this true on the rotating earth assuming the experiment takes place over a large perfectly flat field so the curvature of the earth is not an issue? Explain. What other irregularities will occur? Recall the Coriolis force is \( 2\omega \left[ (-y' \cos \phi) \mathbf{i} + (x' \cos \phi + z' \sin \phi) \mathbf{j} - (y' \sin \phi) \mathbf{k} \right] \) where \( \mathbf{k} \) points away from the center of the earth, \( \mathbf{j} \) points East, and \( \mathbf{i} \) points South.

18. Suppose you have \( n \) masses, \( m_1, \ldots, m_n \). Let the position vector of the \( i^{th} \) mass be \( \mathbf{r}_i(t) \). The center of mass of these is defined to be

\[
\mathbf{R}(t) = \frac{\sum_{i=1}^{n} m_i \mathbf{r}_i(t)}{\sum_{i=1}^{n} m_i}.
\]

Let \( \mathbf{r}_{Bi}(t) = \mathbf{r}_i(t) - \mathbf{R}(t) \). Show that \( \sum_{i=1}^{n} m_i \mathbf{r}_i(t) - \sum_{i=1}^{n} m_i \mathbf{R}(t) = 0 \).

19. Suppose you have \( n \) masses, \( m_1, \ldots, m_n \) which make up a moving rigid body. Let \( \mathbf{R}(t) \) denote the position vector of the center of mass of these \( n \) masses. Find a formula for the total kinetic energy in terms of this position vector, the angular velocity vector, and the position vector of each mass from the center of mass. **Hint:** Use Problem 18.
13.9 Line Integrals

The concept of the integral can be extended to functions which are not defined on an interval of the real line but on some curve in $\mathbb{R}^n$. This is done by defining things in such a way that the more general concept reduces to the earlier notion. First it is necessary to consider what is meant by arc length.

13.9.1 Arc Length And Orientations

The application of the integral considered here is the concept of the length of a curve. $C$ is a smooth curve in $\mathbb{R}^n$ if there exists an interval, $[a, b] \subseteq \mathbb{R}$ and functions $x_i : [a, b] \rightarrow \mathbb{R}$ such that the following conditions hold

1. $x_i$ is continuous on $[a, b]$.
2. $x_i'(a)$ exists and is continuous and bounded on $[a, b]$, with $x_i'(a)$ defined as the derivative from the right, 
   \[ \lim_{h \to 0^+} \frac{x_i(a + h) - x_i(a)}{h}, \]
   and $x_i'(b)$ defined similarly as the derivative from the left.
3. For $p(t) \equiv (x_1(t), \cdots, x_n(t))$, $t \rightarrow p(t)$ is one to one on $(a, b)$.
4. $|p'(t)| \equiv (\sum_{i=1}^{n} |x_i'(t)|^2)^{1/2} \neq 0$ for all $t \in [a, b]$.
5. $C = \cup \{(x_1(t), \cdots, x_n(t)) : t \in [a, b]\}.$

The functions, $x_i(t)$, defined above are giving the coordinates of a point in $\mathbb{R}^n$ and the list of these functions is called a parameterization for the smooth curve. Note the natural direction of the interval also gives a direction for moving along the curve. Such a direction is called an orientation. The integral is used to define what is meant by the length of such a smooth curve. Consider such a smooth curve having parameterization $(x_1, \cdots, x_n)$ . Forming a partition of $[a, b]$, $a = t_0 < \cdots < t_n = b$ and letting $p_i = (x_1(t_i), \cdots, x_n(t_i))$, you could consider the polygon formed by lines from $p_0$ to $p_1$ and from $p_1$ to $p_2$ and from $p_3$ to $p_4$ etc. to be an approximation to the curve, $C$. The following picture illustrates what is meant by this.

![Diagram of a curve with points labeled P0, P1, P2, P3]
Thus the length of the curve is approximated by

$$\sum_{k=1}^{n} |p(t_k) - p(t_{k-1})|.$$

Since the functions in the parameterization are differentiable, it is reasonable to expect this to be close to

$$\sum_{k=1}^{n} |p'(t_{k-1})| (t_k - t_{k-1})$$

which is seen to be a Riemann sum for the integral

$$\int_{a}^{b} |p'(t)| \, dt$$

and it is this integral which is defined as the length of the curve.

Would the same length be obtained if another parameterization were used? This is a very important question because the length of the curve should depend only on the curve itself and not on the method used to trace out the curve. The answer to this question is that the length of the curve does not depend on parameterization. The proof is somewhat technical so is given in the last section of this chapter.

Does the definition of length given above correspond to the usual definition of length in the case when the curve is a line segment? It is easy to see that it does so by considering two points in $\mathbb{R}^n$, $p$ and $q$. A parameterization for the line segment joining these two points is

$$f_i(t) \equiv tp_i + (1-t)q_i, \ t \in [0,1].$$

Using the definition of length of a smooth curve just given, the length according to this definition is

$$\int_{0}^{1} \left( \sum_{i=1}^{n} (p_i - q_i)^2 \right)^{1/2} \, dt = |p - q|.$$

Thus this new definition which is valid for smooth curves which may not be straight line segments gives the usual length for straight line segments.

The proof that curve length is well defined for a smooth curve contains a result which deserves to be stated as a corollary. It is proved in Lemma 14.3.13 on Page 274 but the proof is mathematically fairly advanced so it is presented later.

**Corollary 13.9.1** Let $C$ be a smooth curve and let $f : [a,b] \to C$ and $g : [c,d] \to C$ be two parameterizations satisfying 1 - 5. Then $g^{-1} \circ f$ is either strictly increasing or strictly decreasing.
Definition 13.9.2 If $g^{-1} \circ f$ is increasing, then $f$ and $g$ are said to be equivalent parameterizations and this is written as $f \sim g$. It is also said that the two parameterizations give the same orientation for the curve when $f \sim g$.

When the parameterizations are equivalent, they preserve the direction of motion along the curve and this also shows there are exactly two orientations of the curve since either $g^{-1} \circ f$ is increasing or it is decreasing. This is not hard to believe. In simple language, the message is that there are exactly two directions of motion along a curve. The difficulty is in proving this is actually the case.

Lemma 13.9.3 The following hold for $\sim$.

\begin{align*}
    f & \sim f, \quad (13.23) \\
    \text{If } f \sim g \text{ then } g & \sim f, \quad (13.24) \\
    \text{If } f \sim g \text{ and } g \sim h, \text{ then } f & \sim h. \quad (13.25)
\end{align*}

Proof: Formula (13.23) is obvious because $f^{-1} \circ f (t) = t$ so it is clearly an increasing function. If $f \sim g$ then $f^{-1} \circ g$ is increasing. Now $g^{-1} \circ f$ must also be increasing because it is the inverse of $f^{-1} \circ g$. This verifies (13.24). To see (13.25), $f^{-1} \circ h = (f^{-1} \circ g) \circ (g^{-1} \circ h)$ and so since both of these functions are increasing, it follows $f^{-1} \circ h$ is also increasing. This proves the lemma.

The symbol, $\sim$, is called an equivalence relation. If $C$ is such a smooth curve just described, and if $f : [a, b] \to C$ is a parameterization of $C$, consider $g : (t) \equiv f ((a + b) - t)$, also a parameterization of $C$. Now by Corollary 13.9.1, if $h$ is a parameterization, then if $f^{-1} \circ h$ is not increasing, it must be the case that $g^{-1} \circ h$ is increasing. Consequently, either $h \sim g$ or $h \sim f$. These parameterizations, $h$, which satisfy $h \sim f$ are called the equivalence class determined by $f$ and those $h \sim g$ are called the equivalence class determined by $g$. These two classes are called orientations of $C$. They give the direction of motion on $C$. You see that going from $f$ to $g$ corresponds to tracing out the curve in the opposite direction.

Sometimes people wonder why it is required, in the definition of a smooth curve that $p'(t) \neq 0$. Imagine $t$ is time and $p(t)$ gives the location of a point in space. If $p'(t)$ is allowed to equal zero, the point can stop and change directions abruptly, producing a pointy place in $C$. Here is an example.

Example 13.9.4 Graph the curve $(t^3, t^2)$ for $t \in [-1, 1]$.

In this case, $t = x^{1/3}$ and so $y = x^{2/3}$. Thus the graph of this curve looks like the picture below. Note the pointy place. Such a curve should not be considered smooth.

13.9.2 Line Integrals And Work

Let $C$ be a smooth curve contained in $\mathbb{R}^p$. A curve, $C$, is an “oriented curve” if the only parameterizations considered are those which lie in exactly one of the two equivalence classes, each of which is called an “orientation”. In simple language, orientation specifies a direction over which motion along the curve is to take place. Thus, it specifies the order in which the points of $C$ are encountered. The pair of concepts consisting of the set of points making up the curve along with a direction of motion along the curve is called an oriented curve.
Definition 13.9.5 Suppose $F(x) \in \mathbb{R}^p$ is given for each $x \in C$ where $C$ is a smooth oriented curve and suppose $x \rightarrow F(x)$ is continuous. The mapping $x \rightarrow F(x)$ is called a vector field. In the case that $F(x)$ is a force, it is called a force field.

Next the concept of work done by a force field, $F$ on an object as it moves along the curve, $C$, in the direction determined by the given orientation of the curve will be defined. This is new. Earlier the work done by a force which acts on an object moving in a straight line was discussed but here the object moves over a curve. In order to define what is meant by the work, consider the following picture.

In this picture, the work done by a constant force, $F$ on an object which moves from the point $x(t)$ to the point $x(t+h)$ along the straight line shown would equal $F(x(t+h) - x(t))$. It is reasonable to assume this would be a good approximation to the work done in moving along the curve joining $x(t)$ and $x(t+h)$ provided $h$ is small enough. Also, provided $h$ is small,

$$x(t+h) - x(t) \approx x'(t)h$$

where the wriggly equal sign indicates the two quantities are close. In the notation of Leibniz, one writes $dt$ for $h$ and

$$dW = F(x(t)) \cdot x'(t)dt$$

or in other words,

$$\frac{dW}{dt} = F(x(t)) \cdot x'(t).$$

Defining the total work done by the force at $t = 0$, corresponding to the first endpoint of the curve, to equal zero, the work would satisfy the following initial value problem.

$$\frac{dW}{dt} = F(x(t)) \cdot x'(t), \quad W(a) = 0.$$ 

This motivates the following definition of work.

Definition 13.9.6 Let $F(x)$ be given above. Then the work done by this force field on an object moving over the curve $C$ in the direction determined by the specified orientation is defined as

$$\int_C F \cdot dR \equiv \int_a^b F(x(t)) \cdot x'(t) \, dt$$

where the function, $x$ is one of the allowed parameterizations of $C$ in the given orientation of $C$. In other words, there is an interval, $[a,b]$ and as $t$ goes from $a$ to $b$, $x(t)$ moves in the direction determined from the given orientation of the curve.
**Theorem 13.9.7** The symbol, \( \int_C \mathbf{F} \cdot d\mathbf{R} \), is well defined in the sense that every parameterization in the given orientation of \( C \) gives the same value for \( \int_C \mathbf{F} \cdot d\mathbf{R} \).

**Proof:** Suppose \( g : [c, d] \to C \) is another allowed parameterization. Thus \( g^{-1} \circ f \) is an increasing function, \( \phi \). Then since \( \phi \) is increasing,

\[
\int_c^d F(g(s)) \cdot g'(s) \, ds = \int_a^b F(g(\phi(t))) \cdot g'(\phi(t)) \phi'(t) \, dt
\]

This proves the theorem.

Regardless the physical interpretation of \( F \), this is called the line integral. When \( F \) is interpreted as a force, the line integral measures the extent to which the motion over the curve in the indicated direction is aided by the force. If the net effect of the force on the object is to impede rather than to aid the motion, this will show up as the work being negative.

Does the concept of work as defined here coincide with the earlier concept of work when the object moves over a straight line when acted on by a constant force?

Let \( p \) and \( q \) be two points in \( \mathbb{R}^n \) and suppose \( F \) is a constant force acting on an object which moves from \( p \) to \( q \) along the straight line joining these points. Then the work done is \( F \cdot (q - p) \). Is the same thing obtained from the above definition? Let \( x(t) \equiv p + t(q - p), t \in [0, 1] \) be a parameterization for this oriented curve, the straight line in the direction from \( p \) to \( q \). Then \( x'(t) = q - p \) and \( F(x(t)) = F \). Therefore, the above definition yields

\[
\int_0^1 F \cdot (q - p) \, dt = F \cdot (q - p)
\]

Therefore, the new definition adds to but does not contradict the old one.

**Example 13.9.8** Suppose for \( t \in [0, \pi] \) the position of an object is given by \( r(t) = t\mathbf{i} + \cos(2t)\mathbf{j} + \sin(2t)\mathbf{k} \). Also suppose there is a force field defined on \( \mathbb{R}^3 \), \( F(x, y, z) \equiv 2xy\mathbf{i} + x^2\mathbf{j} + \mathbf{k} \). Find

\[
\int_C F \cdot d\mathbf{R}
\]

where \( C \) is the curve traced out by this object which has the orientation determined by the direction of increasing \( t \).

To find this line integral use the above definition and write

\[
\int_C F \cdot d\mathbf{R} = \int_0^\pi (2t \cos(2t)) \cdot (1, -2\sin(2t), 2\cos(2t)) \, dt
\]

In evaluating this replace the \( x \) in the formula for \( F \) with \( t \), the \( y \) in the formula for \( F \) with \( \cos(2t) \) and the \( z \) in the formula for \( F \) with \( \sin(2t) \) because these are the values of these variables which correspond to the value of \( t \). Taking the dot product, this equals the following integral.

\[
\int_0^\pi (2t \cos 2t - 2 (\sin 2t) t^2 + 2 \cos 2t) \, dt = \pi^2
\]
Example 13.9.9 Let $C$ denote the oriented curve obtained by $r(t) = (t, \sin t, t^3)$ where the orientation is determined by increasing $t$ for $t \in [0, 2]$. Also let $F = (x, y, xz + z)$. Find $\int_C F \cdot dR$.

You use the definition.

$$\int_C F \cdot dR = \int_0^2 (t, \sin t, (t + 1)^3) \cdot (1, \cos t, 3t^2) \, dt$$

$$= \int_0^2 (t + \sin t \cos t + 3 \cdot (t + 1)^3) \, dt$$

$$= \frac{1251}{14} - \frac{1}{2} \cos^2 (2).$$

Suppose you have a curve specified by $r(s) = (x(s), y(s), z(s))$ and it has the property that $|r'(s)| = 1$ for all $s \in [0, b]$. Then the length of this curve for $s$ between $0$ and $s_1$ is

$$\int_0^{s_1} |r'(s)| \, ds = \int_0^{s_1} 1 \, ds = s_1.$$  

This parameter is therefore called arc length because the length of the curve up to $s$ equals $s$. Now you can always change the parameter to be arc length.

Proposition 13.9.10 Suppose $C$ is an oriented smooth curve parameterized by $r(t)$ for $t \in [a, b]$. Then letting $l$ denote the total length of $C$, there exists $R(s), s \in [0, l]$ another parameterization for this curve which preserves the orientation and such that $|R'(s)| = 1$ so that $s$ is arc length.

Prove: Let $\phi(t) \equiv \int_a^t |r'(\tau)| \, d\tau \equiv s$. Then $s$ is an increasing function of $t$ because

$$\frac{ds}{dt} = \phi'(t) = |r'(t)| > 0.$$  

Now define $R(s) \equiv r(\phi^{-1}(s))$. Then

$$R'(s) = \frac{r'(\phi^{-1}(s))}{|r'(\phi^{-1}(s))|}$$

and so $|R'(s)| = 1$ as claimed. $R(l) = r(\phi^{-1}(l)) = r(\phi^{-1}\left(\int_a^b |r'(\tau)| \, d\tau\right)) = r(b)$ and $R(0) = r(\phi^{-1}(0)) = r(a)$ and $R$ delivers the same set of points in the same order as $r$ because $\frac{ds}{dt} > 0$.

The arc length parameter is just like any other parameter in so far as considerations of line integrals are concerned because it was shown above that line integrals are independent of parameterization. However, when things are defined in terms of the arc length parameterization, it is clear they depend only on geometric properties of the curve itself and for this reason, the arc length parameterization is important in differential geometry.

13.9.3 Another Notation For Line Integrals

Definition 13.9.11 Let $F(x, y, z) = (P(x, y, z), Q(x, y, z), R(x, y, z))$ and let $C$ be an oriented curve. Then another way to write $\int_C F \cdot dR$ is

$$\int_C P \, dx + Q \, dy + R \, dz.$$
This last is referred to as the integral of a differential form, \( Pdx + Qdy + Rdz \). The study of differential forms is important. Formally, \( d\mathbf{R} = (dx, dy, dz) \) and so the integrand in the above is formally \( \mathbf{F} \cdot d\mathbf{R} \). Other occurrences of this notation are handled similarly in 2 or higher dimensions.

### 13.10 Exercises

1. Suppose for \( t \in [0, 2\pi] \) the position of an object is given by \( \mathbf{r}(t) = ti + \cos(2t)j + \sin(2t)k \). Also suppose there is a force defined on \( \mathbb{R}^3 \), \( \mathbf{F}(x, y, z) \equiv 2xzi + (x^2 + 2yz)j + y^2k \). Find the work,

\[
\int_C \mathbf{F} \cdot d\mathbf{R}
\]

where \( C \) is the curve traced out by this object which has the orientation determined by the direction of increasing \( t \).

2. Here is a vector field, \((y, x + z^2, 2yz)\) and here is the parameterization of a curve, \( \mathbf{R}(t) = (\cos 2t, 2\sin 2t, t) \) where \( t \) goes from 0 to \( \pi/4 \). Find \( \int_C \mathbf{F} \cdot d\mathbf{R} \).

3. If \( f \) and \( g \) are both increasing functions, show \( f \circ g \) is an increasing function also. Assume anything you like about the domains of the functions.

4. Suppose for \( t \in [0,3] \) the position of an object is given by \( \mathbf{r}(t) = ti + tj + tk \). Also suppose there is a force defined on \( \mathbb{R}^3 \), \( \mathbf{F}(x, y, z) \equiv yzi + xzj + xyk \). Find

\[
\int_C \mathbf{F} \cdot d\mathbf{R}
\]

where \( C \) is the curve traced out by this object which has the orientation determined by the direction of increasing \( t \). Repeat the problem for \( \mathbf{r}(t) = ti + t^2j + tk \).

5. Suppose for \( t \in [0,1] \) the position of an object is given by \( \mathbf{r}(t) = ti + tj + tk \). Also suppose there is a force defined on \( \mathbb{R}^3 \), \( \mathbf{F}(x, y, z) \equiv zi + xzj + xyk \). Find

\[
\int_C \mathbf{F} \cdot d\mathbf{R}
\]

where \( C \) is the curve traced out by this object which has the orientation determined by the direction of increasing \( t \). Repeat the problem for \( \mathbf{r}(t) = ti + t^2j + tk \).

6. Let \( \mathbf{F}(x, y, z) \) be a given force field and suppose it acts on an object having mass, \( m \) on a curve with parameterization, \((x(t), y(t), z(t))\) for \( t \in [a, b] \). Show directly that the work done equals the difference in the kinetic energy. **Hint:**

\[
\int_a^b \mathbf{F}(x(t), y(t), z(t)) \cdot (x'(t), y'(t), z'(t)) \, dt = m \int_a^b \mathbf{r}(t) \cdot \mathbf{r}'(t) \, dt,
\]

etc.
14.0.1 Outcomes

1. Recall the definitions of unit tangent, unit normal, and osculating plane.

2. Calculate the curvature for a space curve.

3. Given the position vector function of a moving object, calculate the velocity, speed, and acceleration of the object and write the acceleration in terms of its tangential and normal components.

4. Derive formulas for the curvature of a parameterized curve and the curvature of a plane curve given as a function.

14.1 Space Curves

A fly buzzing around the room, a person riding a roller coaster, and a satellite orbiting the earth all have something in common. They are moving over some sort of curve in three dimensions.

Denote by \( R(t) \) the function which takes \( t \) to a point on this curve where \( t \) is time. Thus \( R(t) \) equals the point on the curve which occurs at time \( t \). Assume that \( R', R'' \) exist and is continuous. Thus \( R' = v \), the velocity and \( R'' = a \) is the acceleration.

**Lemma 14.1.1** Define \( T(t) \equiv \frac{R'(t)}{|R'(t)|} \). Then \( |T(t)| = 1 \) and if \( T'(t) \neq 0 \), then there exists a unit vector, \( N(t) \) called the principle normal perpendicular to \( T(t) \) and a scalar valued function, \( \kappa(t) \) with \( T'(t) = \kappa(t)|v|N(t) \).

**Proof:** It follows from the definition that \(|T| = 1\). Therefore, \( T \cdot T = 1 \) and so, upon differentiating both sides,

\[ T' \cdot T + T \cdot T' = 2T' \cdot T = 0. \]

Therefore, \( T' \) is perpendicular to \( T \). Let

\[ N(t) \equiv \frac{T'}{|T'|}. \]

Then letting \(|T'| \equiv \kappa(t)|v(t)| \), it follows

\[ T'(t) = \kappa(t)|v(t)|N(t). \]

This proves the lemma.
The plane determined by the two vectors, $T$ and $N$ is called the osculating plane. It identifies a particular plane which is in a sense tangent to this space curve. In the case where $|T'(t)| = 0$ near the point of interest, $T(t)$ equals a constant and so the space curve is a straight line which it would be supposed has no curvature. Also, the principal normal is undefined in this case. This makes sense because if there is no curving going on, there is no special direction normal to the curve at such points which could be distinguished from any other direction normal to the curve. In the case where $|T'(t)| = 0$, $\kappa(t) = 0$ and the radius of curvature would be considered infinite.

**Definition 14.1.2** The vector, $T(t)$ is called the unit tangent vector and the vector, $N(t)$ is called the principal normal. The function, $\kappa(t)$ in the above lemma is called the curvature. The radius of curvature is defined as $\rho = \frac{1}{\kappa}$.

The important thing about this is that it is possible to write the acceleration as the sum of two vectors, one perpendicular to the direction of motion and the other in the direction of motion.

**Theorem 14.1.3** For $R(t)$ the position vector of a space curve, the acceleration is given by the formula

$$a = \frac{d|v|}{dt}T + \kappa|v|^2N$$

(14.1)

Furthermore, $a_T^2 + a_N^2 = |a|^2$.

**Proof:**

$$a = \frac{dv}{dt} = \frac{d}{dt}(R') = \frac{d}{dt}(|v|T) = \frac{d|v|}{dt}T + |v|T' = \frac{d}{dt}|v|^2T$$

This proves the first part.

For the second part,

$$|a|^2 = (a_TT + a_NN) \cdot (a_TT + a_NN) = a_T^2T \cdot T + 2a_Na_TT \cdot N + a_N^2N \cdot N = a_T^2 + a_N^2$$

because $T \cdot N = 0$. This proves the theorem.

Finally, it is well to point out that the curvature is a property of the curve itself, and does not depend on the parameterization of the curve. If the curve is given by two different vector valued functions, $R(t)$ and $R(\tau)$, then from the formula above for the curvature,

$$\kappa(t) = \frac{|T'(t)|}{|v(t)|} = \frac{|dT}{d\tau} \frac{d\tau}{dt} = \frac{|dT}{d\tau} \frac{d\tau}{d\tau} \equiv \kappa(\tau).$$

From this, it is possible to give an important formula from physics. Suppose an object orbits a point at constant speed, $v$. In the above notation, $|v| = v$. What is the centripetal

---

1To osculate means to kiss. Thus this plane could be called the kissing plane. However, that does not sound formal enough so we call it the osculating plane.
acceleration of this object? You may know from a physics class that the answer is \(v^2/r\) where \(r\) is the radius. This follows from the above quite easily. The parameterization of the object which is as described is

\[ \mathbf{R}(t) = \left( r \cos \left( \frac{v}{r} t \right), r \sin \left( \frac{v}{r} t \right) \right). \]

Therefore, \(\mathbf{T} = (-\sin \left( \frac{v}{r} t \right), \cos \left( \frac{v}{r} t \right))\) and \(\mathbf{T}' = (-\frac{v}{r} \cos \left( \frac{v}{r} t \right), -\frac{v}{r} \sin \left( \frac{v}{r} t \right))\). Thus, \(\kappa = |\mathbf{T}'(t)|/v = \frac{1}{r}\). It follows

\[ \mathbf{a} = \frac{dv}{dt} \mathbf{T} + v^2 \kappa \mathbf{N} = \frac{v^2}{r} \mathbf{N}. \]

The vector, \(\mathbf{N}\) points from the object toward the center of the circle because it is a positive multiple of the vector, \((-\frac{v}{r} \cos \left( \frac{v}{r} t \right), -\frac{v}{r} \sin \left( \frac{v}{r} t \right))\).

Formula (14.1) also yields an easy way to find the curvature. Take the cross product of both sides with \(\mathbf{v}\), the velocity. Then

\[ \mathbf{a} \times \mathbf{v} = \frac{d|\mathbf{v}|}{dt} \mathbf{T} \times \mathbf{v} + |\mathbf{v}|^2 \kappa \mathbf{N} \times \mathbf{v} \]

\[ = \frac{d|\mathbf{v}|}{dt} \mathbf{T} \times \mathbf{v} + |\mathbf{v}|^3 \kappa \mathbf{N} \times \mathbf{T} \]

Now \(\mathbf{T}\) and \(\mathbf{v}\) have the same direction so the first term on the right equals zero. Taking the magnitude of both sides, and using the fact that \(\mathbf{N}\) and \(\mathbf{T}\) are two perpendicular unit vectors,

\[ |\mathbf{a} \times \mathbf{v}| = |\mathbf{v}|^3 \kappa \]

and so

\[ \kappa = \frac{|\mathbf{a} \times \mathbf{v}|}{|\mathbf{v}|^3}. \quad (14.2) \]

**Example 14.1.4** Let \(\mathbf{R}(t) = (\cos (t), t, t^2)\) for \(t \in [0, 3]\). Find the speed, velocity, curvature, and write the acceleration in terms of normal and tangential components.

First of all \(\mathbf{v}(t) = (-\sin t, 1, 2t)\) and so the speed is given by

\[ |\mathbf{v}| = \sqrt{\sin^2(t) + 1 + 4t^2}. \]

Therefore,

\[ a_T = \frac{d}{dt} \left( \sqrt{\sin^2(t) + 1 + 4t^2} \right) = \frac{\sin(t) \cos(t) + 4t}{\sqrt{(2 + 4t^2 - \cos^2(t)}} \]

It remains to find \(a_N\). To do this, you can find the curvature first if you like.

\[ \mathbf{a}(t) = \mathbf{R}''(t) = (-\cos t, 0, 2). \]

Then

\[ \kappa = \frac{|(-\cos t, 0, 2) \times (-\sin t, 1, 2t)|}{\left( \sqrt{\sin^2(t) + 1 + 4t^2} \right)^3} \]

\[ = \frac{\sqrt{4 + (-2 \sin(t) + 2 (\cos(t)))^2 + \cos^2(t)}}{\left( \sqrt{\sin^2(t) + 1 + 4t^2} \right)^3} \]
Then
\[ a_N = \kappa |v|^2 \]
\[ = \sqrt{4 + (-2 \sin (t) + 2 (\cos (t))t)^2 + \cos^2 (t)} \left( 3 \sin^2 (t) + 1 + 4t^2 \right) \]
\[ = \frac{\sqrt{4 + (-2 \sin (t) + 2 (\cos (t))t)^2 + \cos^2 (t)}}{\sin^2 (t) + 1 + 4t^2} \cdot \sqrt{\sin^2 (t) + 1 + 4t^2} \].

You can observe the formula \(a_N^2 + a_T^2 = |\mathbf{a}|^2\) holds. Indeed \(a_N^2 + a_T^2 = \)
\[ \left( \frac{\sqrt{4 + (-2 \sin (t) + 2 (\cos (t))t)^2 + \cos^2 (t)}}{\sin^2 (t) + 1 + 4t^2} \right)^2 + \left( \frac{\sin (t) \cos (t) + 4t}{\sqrt{2 + 4t^2 - \cos^2 (t)}} \right)^2 \]
\[ = \frac{4 + (-2 \sin t + 2 (\cos t)t)^2 + \cos^2 t}{\sin^2 t + 1 + 4t^2} + \frac{(\sin t \cos t + 4t)^2}{2 + 4t^2 - \cos^2 t} = \cos^2 t + 4 = |\mathbf{a}|^2 \]

**Example 14.1.5** Find a formula for the curvature of the curve given by the graph of \(y = f(x)\) for \(x \in [a, b]\). Assume whatever you like about smoothness of \(f\).

You need to write this as a parametric curve. This is most easily accomplished by letting \(t = x\). Thus a parameterization is
\[(t, f(t), 0) : t \in [a, b].\]

Then you can use the formula given above. The acceleration is \((0, f''(t), 0)\) and the velocity is \((1, f'(t), 0)\). Therefore,
\[ \mathbf{a} \times \mathbf{v} = (0, f''(t), 0) \times (1, f'(t), 0) = (0, 0, -f''(t)). \]

Therefore, the curvature is given by
\[ \frac{|\mathbf{a} \times \mathbf{v}|}{|\mathbf{v}|^2} = \frac{|f''(t)|}{\left(1 + f'(t)^2\right)^{3/2}}. \]

Sometimes curves don’t come to you parametrically. This is unfortunate when it occurs but you can sometimes find a parametric description of such curves. It should be emphasized that it is only sometimes when you can actually find a parameterization. General systems of nonlinear equations cannot be solved using algebra.

**Example 14.1.6** Find a parameterization for the intersection of the surfaces \(y + 3z = 2x^2 + 4\) and \(y + 2z = x + 1\).

You need to solve for \(x\) and \(y\) in terms of \(x\). This yields
\[ z = 2x^2 - x + 3, \ y = -4x^2 + 3x - 5. \]

Therefore, letting \(t = x\), the parameterization is \((x, y, z) = (t, -4t^2 - 5 + 3t, -t + 3 + 2t^2)\).

**Example 14.1.7** Find a parameterization for the straight line joining \((3, 2, 4)\) and \((1, 10, 5)\).

\((x, y, z) = (3, 2, 4) + t(-2, 8, 1) = (3 - 2t, 2 + 8t, 4 + t)\) where \(t \in [0, 1]\). Note where this came from. The vector, \((-2, 8, 1)\) is obtained from \((1, 10, 5) - (3, 2, 4)\). Now you should check to see this works.
14.2 Exercises

1. Find a parametrization for the intersection of the planes $2x + y + 3z = -2$ and $3x - 2y + z = -4$.

2. Find a parametrization for the intersection of the plane $3x + y + z = -3$ and the circular cylinder $x^2 + y^2 = 1$.

3. Find a parametrization for the intersection of the plane $4x + 2y + 3z = 2$ and the elliptic cylinder $x^2 + 4z^2 = 9$.

4. Find a parametrization for the straight line joining $(1, 2, 1)$ and $(-1, 4, 4)$.

5. Find a parametrization for the intersection of the surfaces $3y + 3z = 3x^2$ and $3y + 2z = 3$.

6. An object moves in $\mathbb{R}^3$ according to the formula, $(\cos 3t, \sin 3t, t^2)$. Find the normal and tangential components of the acceleration of this object as a function of $t$ and write the acceleration in the form $a_T T + a_N N$.

7. An object moves over the circle of radius $r$ according to the formula, $(r \cos (\omega t), r \sin (\omega t))$ where $v = r \omega$. Show that the speed of the object is constant and equals to $v$. Tell why $a_T = 0$ and find $a_N$, $N$. This yields the formula for centripetal acceleration from beginning physics classes.

8. Suppose $|\mathbf{R}(t)| = c$ where $c$ is a constant and $\mathbf{R}(t)$ is the position vector of an object. Show the velocity, $\mathbf{R}'(t)$ is always perpendicular to $\mathbf{R}(t)$.

9. An object moves in the plane according to the formula, $(\cos 3t, \sin 3t, t^2)$. Find the normal and tangential components of the acceleration of this object as a function of $t$ and write the acceleration in the form $a_T T + a_N N$.

10. Find a formula for the curvature of the curve, $y = \sin x$ in the $xy$ plane.

11. A particle moves in the plane according to the formula, $(\cos 3t, \sin 3t, t^2)$. Find the normal and tangential components of the acceleration of this object as a function of $t$ and write the acceleration in the form $a_T T + a_N N$.

12. An object moves over a helix, $(\cos t, \sin t, t)$.

13. An object moves over the helix, $(\cos t, \sin t, t^2)$.

14. An object moves in three dimensions and the only force on the object is a central force. This means that if $\mathbf{r}(t)$ is the position of the object, $\mathbf{a}(t) = k(\mathbf{r}(t)) \mathbf{r}(t)$ where $k$ is some function. Show that if this happens, then the motion of the object must be in a plane. **Hint:** First argue that $\mathbf{a} \times \mathbf{r} = 0$. Next show that $(\mathbf{a} \times \mathbf{r}) = (\mathbf{v} \times \mathbf{r})'$. Therefore, $(\mathbf{v} \times \mathbf{r})' = 0$. Explain why this requires $\mathbf{v} \times \mathbf{r} = \mathbf{c}$ for some vector, $\mathbf{c}$ which does not depend on $t$. Then explain why $\mathbf{c} \cdot \mathbf{r} = 0$. This implies the motion is in a plane. Why? What are some examples of central forces?

15. Let $\mathbf{R}(t) = (\cos t) \mathbf{i} + (\cos t) \mathbf{j} + (\sqrt{2} \sin t) \mathbf{k}$. Find the arc length, $s$ as a function of the parameter, $t$, if $t = 0$ is taken to correspond to $s = 0$. 
16. Let $\mathbf{R}(t) = 2\mathbf{i} + (4t + 2)\mathbf{j} + 4t\mathbf{k}$. Find the arc length, $s$ as a function of the parameter, $t$, if $t = 0$ is taken to correspond to $s = 0$.

17. Let $\mathbf{R}(t) = e^{5t}\mathbf{i} + e^{-5t}\mathbf{j} + 5\sqrt{2}t\mathbf{k}$. Find the arc length, $s$ as a function of the parameter, $t$, if $t = 0$ is taken to correspond to $s = 0$.

18. An object moves along the $x$ axis toward $(0,0)$ and then along the curve $y = x^2$ in the direction of increasing $x$ at constant speed. Is the force acting on the object a continuous function? Explain. Is there any physically reasonable way to make this force continuous by relaxing the requirement that the object move at constant speed? If the curve were part of a railroad track, what would happen at the point where $x = 0$?

19. An object of mass $m$ moving over a space curve is acted on by a force, $\mathbf{F}$. Show the work done by this force equals $ma_T$ (length of the curve). In other words, it is only the tangential component of the force which does work.

### 14.3 Independence Of Parameterization*

Recall that if $p(t) : t \in [a, b]$ was a parameterization of a smooth curve, $C$, the length of $C$ is defined as

$$\int_a^b |p'(t)| \, dt$$

If some other parameterization were used to trace out $C$, would the same answer be obtained? To answer this question in a satisfactory manner requires some hard calculus.

#### 14.3.1 Hard Calculus

**Definition 14.3.1** A sequence $\{a_n\}_{n=1}^\infty$ converges to $a$,

$$\lim_{n \to \infty} a_n = a \text{ or } a_n \to a$$

if and only if for every $\varepsilon > 0$ there exists $n_\varepsilon$ such that whenever $n \geq n_\varepsilon$,

$$|a_n - a| < \varepsilon.$$

In words the definition says that given any measure of closeness, $\varepsilon$, the terms of the sequence are eventually all this close to $a$. Note the similarity with the concept of limit. Here, the word “eventually” refers to $n$ being sufficiently large. Earlier, it referred to $y$ being sufficiently close to $x$ on one side or another or else $x$ being sufficiently large in either the positive or negative directions. The limit of a sequence, if it exists, is unique.
Theorem 14.3.2 If \( \lim_{n \to \infty} a_n = a \) and \( \lim_{n \to \infty} a_n = a_1 \) then \( a_1 = a \).

**Proof:** Suppose \( a_1 \neq a \). Then let \( 0 < \varepsilon < \frac{|a_1 - a|}{2} \) in the definition of the limit. It follows there exists \( n_\varepsilon \) such that if \( n \geq n_\varepsilon \), then \( |a_n - a| < \varepsilon \) and \( |a_n - a_1| < \varepsilon \). Therefore, for such \( n \),

\[
|a_1 - a| \leq |a_1 - a_n| + |a_n - a| < \varepsilon + \varepsilon < |a_1 - a|/2 + |a_1 - a|/2 = |a_1 - a|,
\]

a contradiction.

**Definition 14.3.3** Let \( \{a_n\} \) be a sequence and let \( n_1 < n_2 < n_3, \ldots \) be any strictly increasing list of integers such that \( n_1 \) is at least as large as the first index used to define the sequence \( \{a_n\} \). Then if \( b_k \equiv a_{n_k} \), \( \{b_k\} \) is called a subsequence of \( \{a_n\} \).

**Theorem 14.3.4** Let \( \{x_n\} \) be a sequence with \( \lim_{n \to \infty} x_n = x \) and let \( \{x_{n_k}\} \) be a subsequence. Then \( \lim_{k \to \infty} x_{n_k} = x \).

**Proof:** Let \( \varepsilon > 0 \) be given. Then there exists \( n_\varepsilon \) such that if \( n > n_\varepsilon \), then \( |x_n - x| < \varepsilon \). Suppose \( k > n_\varepsilon \). Then \( n_k \geq k > n_\varepsilon \) and so

\[
|x_{n_k} - x| < \varepsilon
\]

showing \( \lim_{k \to \infty} x_{n_k} = x \) as claimed.

There is a very useful way of thinking of continuity in terms of limits of sequences found in the following theorem. In words, it says a function is continuous if it takes convergent sequences to convergent sequences whenever possible.

**Theorem 14.3.5** A function \( f : D(f) \to \mathbb{R} \) is continuous at \( x \in D(f) \) if and only if, whenever \( x_n \to x \) with \( x_n \in D(f) \), it follows \( f(x_n) \to f(x) \).

**Proof:** Suppose first that \( f \) is continuous at \( x \) and let \( x_n \to x \). Let \( \varepsilon > 0 \) be given. By continuity, there exists \( \delta > 0 \) such that if \( |y - x| < \delta \), then \( |f(x) - f(y)| < \varepsilon \). However, there exists \( n_\delta \) such that if \( n \geq n_\delta \), then \( |x_n - x| < \delta \) and so for all \( n \) this large,

\[
|f(x) - f(x_n)| < \varepsilon
\]

which shows \( f(x_n) \to f(x) \).

Now suppose the condition about taking convergent sequences to convergent sequences holds at \( x \). Suppose \( f \) fails to be continuous at \( x \). Then there exists \( \varepsilon > 0 \) and \( x_n \in D(f) \) such that \( |x - x_n| < \frac{1}{n^2} \), yet

\[
|f(x) - f(x_n)| \geq \varepsilon.
\]

But this is clearly a contradiction because, although \( x_n \to x \), \( f(x_n) \) fails to converge to \( f(x) \). It follows \( f \) must be continuous after all. This proves the theorem.

**Definition 14.3.6** A set, \( K \subseteq \mathbb{R} \) is sequentially compact if whenever \( \{a_n\} \subseteq K \) is a sequence, there exists a subsequence, \( \{a_{n_k}\} \) such that this subsequence converges to a point of \( K \).

The following theorem is part of a major advanced calculus theorem known as the Heine Borel theorem.

**Theorem 14.3.7** Every closed interval, \([a, b]\) is sequentially compact.
Proof: Let \( \{x_n\} \subseteq [a, b] \equiv I_0. \) Consider the two intervals \([a, a + b/2]\) and \([a + b/2, b]\) each of which has length \((b - a)/2.\) At least one of these intervals contains \(x_n\) for infinitely many values of \(n.\) Call this interval \(I_1.\) Now do for \(I_1\) what was done for \(I_0.\) Split it in half and let \(I_2\) be the interval which contains \(x_n\) for infinitely many values of \(n.\) Continue this way obtaining a sequence of nested intervals \(I_0 \supseteq I_1 \supseteq I_2 \supseteq I_3 \cdots\) where the length of \(I_n\) is \((b - a)/2^n.\) Now pick \(n_1\) such that \(x_{n_1} \in I_1, n_2\) such that \(n_2 > n_1\) and \(x_{n_2} \in I_2, n_3\) such that \(n_3 > n_2\) and \(x_{n_3} \in I_3, \ldots.\) (This can be done because in each case the intervals contained \(x_n\) for infinitely many values of \(n.\)) By the nested interval lemma there exists a point, \(c\) contained in all these intervals. Furthermore, 

\[
|x_{n_k} - c| < (b - a) 2^{-k}
\]

and so \(\lim_{k \to \infty} x_{n_k} = c \in [a, b].\) This proves the theorem.

Lemma 14.3.8 Let \(\phi : [a, b] \to \mathbb{R}\) be a continuous function and suppose \(\phi\) is \(1 - 1\) on \((a, b).\) Then \(\phi\) is either strictly increasing or strictly decreasing on \([a, b].\) Furthermore, \(\phi^{-1}\) is continuous.

Proof: First it is shown that \(\phi\) is either strictly increasing or strictly decreasing on \((a, b).\)

If \(\phi\) is not strictly decreasing on \((a, b),\) then there exists \(x_1 < y_1, x_1, y_1 \in (a, b)\) such that 

\[
(\phi (y_1) - \phi (x_1))(y_1 - x_1) > 0.
\]

If for some other pair of points, \(x_2 < y_2\) with \(x_2, y_2 \in (a, b),\) the above inequality does not hold, then since \(\phi\) is \(1 - 1,\)

\[
(\phi (y_2) - \phi (x_2))(y_2 - x_2) < 0.
\]

Let \(x_t \equiv tx_1 + (1 - t)x_2\) and \(y_t \equiv ty_1 + (1 - t)y_2.\) Then \(x_t < y_t\) for all \(t \in [0, 1]\) because 

\[
tx_1 \leq ty_1 \text{ and } (1 - t)x_2 \leq (1 - t)y_2
\]

with strict inequality holding for at least one of these inequalities since not both \(t\) and \((1 - t)\) can equal zero. Now define 

\[
h (t) \equiv (\phi (y_t) - \phi (x_t))(y_t - x_t).
\]

Since \(h\) is continuous and \(h(0) < 0,\) while \(h(1) > 0,\) there exists \(t \in (0, 1)\) such that 

\[
h (t) = 0.
\]

Therefore, both \(x_t\) and \(y_t\) are points of \((a, b)\) and \(\phi (y_t) - \phi (x_t) = 0\) contradicting the assumption that \(\phi\) is one to one. It follows \(\phi\) is either strictly increasing or strictly decreasing on \((a, b).\)

This property of being either strictly increasing or strictly decreasing on \((a, b)\) carries over to \([a, b]\) by the continuity of \(\phi.\) Suppose \(\phi\) is strictly increasing on \((a, b),\) a similar argument holding for \(\phi\) strictly decreasing on \((a, b).\) If \(x > a,\) then pick \(y \in (a, x)\) and from the above, \(\phi (y) < \phi (x).\) Now by continuity of \(\phi\) at \(a,\)

\[
\phi (a) = \lim_{z \to a^+} \phi (z) \leq \phi (y) < \phi (x).
\]

Therefore, \(\phi (a) < \phi (x)\) whenever \(x \in (a, b).\) Similarly \(\phi (b) > \phi (x)\) for all \(x \in (a, b).\)

It only remains to verify \(\phi^{-1}\) is continuous. Suppose then that \(s_n \to s\) where \(s_n\) and \(s\) are points of \(\phi ([a, b]).\) It is desired to verify that \(\phi^{-1} (s_n) \to \phi^{-1} (s).\) If this does not happen, there exists \(\varepsilon > 0\) and a subsequence, still denoted by \(s_n\) such that \(|\phi^{-1} (s_n) - \phi^{-1} (s)| \geq \varepsilon.\)

Using the sequential compactness of \([a, b]\) there exists a further subsequence, still denoted by \(n,\) such that \(\phi^{-1} (s_n) \to t_1 \in [a, b], t_1 \neq \phi^{-1} (s).\) Then by continuity of \(\phi,\) it follows 

\[
s_n \to \phi (t_1)\]

and so \(s = \phi (t_1).\) Therefore, \(t_1 = \phi^{-1} (s)\) after all. This proves the lemma.
Corollary 14.3.9 Let \( f : (a, b) \to \mathbb{R} \) be one to one and continuous. Then \( f (a, b) \) is an open interval, \((c, d)\) and \( f^{-1} : (c, d) \to (a, b) \) is continuous.

**Proof:** Since \( f \) is either strictly increasing or strictly decreasing, it follows that \( f (a, b) \) is an open interval, \((c, d)\). Assume \( f \) is decreasing. Now let \( x \in (a, b) \). Why is \( f^{-1} \) is continuous at \( f (x) \)? Since \( f \) is decreasing, if \( f (x) < f (y) \), then \( y \equiv f^{-1} (f (y)) < x \equiv f^{-1} (f (x)) \) and so \( f^{-1} \) is also decreasing. Let \( \varepsilon > 0 \) be given. Let \( \varepsilon > \eta > 0 \) and \( (x - \eta, x + \eta) \subseteq (a, b) \). Then \( f (x) \in (f (x + \eta), f (x - \eta)) \). Let \( \delta = \min \left( f (x) - f (x + \eta), f (x - \eta) - f (x) \right) \). Then if
\[
|f (z) - f (x)| < \delta,
\]
it follows
\[
z \equiv f^{-1} (f (z)) \in (x - \eta, x + \eta) \subseteq (x - \varepsilon, x + \varepsilon)
\]
so
\[
|f^{-1} (f (z)) - x| = |f^{-1} (f (z)) - f^{-1} (f (x))| < \varepsilon.
\]
This proves the theorem in the case where \( f \) is strictly decreasing. The case where \( f \) is increasing is similar.

Theorem 14.3.10 Let \( f : [a, b] \to \mathbb{R} \) be continuous and one to one. Suppose \( f' (x_1) \) exists for some \( x_1 \in [a, b] \) and \( f' (x_1) \neq 0 \). Then \( (f^{-1})' (f (x_1)) \) exists and is given by the formula,
\[
(f^{-1})' (f (x_1)) = \frac{1}{f' (x_1)}.
\]

**Proof:** By Lemma 14.3.8 \( f \) is either strictly increasing or strictly decreasing and \( f^{-1} \) is continuous on \([a, b] \). Therefore there exists \( \eta > 0 \) such that if \( 0 < |f (x_1) - f (x)| < \eta \), then
\[
0 < |x_1 - x| = |f^{-1} (f (x_1)) - f^{-1} (f (x))| < \delta
\]
where \( \delta \) is small enough that for \( 0 < |x_1 - x| < \delta \),
\[
\left| \frac{x - x_1}{f (x) - f (x_1)} - \frac{1}{f' (x_1)} \right| < \varepsilon.
\]
It follows that if \( 0 < |f (x_1) - f (x)| < \eta \),
\[
\left| \frac{f^{-1} (f (x)) - f^{-1} (f (x_1))}{f (x) - f (x_1)} - \frac{1}{f' (x_1)} \right| = \left| \frac{x - x_1}{f (x) - f (x_1)} - \frac{1}{f' (x_1)} \right| < \varepsilon
\]
Therefore, since \( \varepsilon > 0 \) is arbitrary,
\[
\lim_{y \to f (x_1)} \frac{f^{-1} (y) - f^{-1} (f (x_1))}{y - f (x_1)} = \frac{1}{f' (x_1)}
\]
and this proves the theorem.

The following obvious corollary comes from the above by not bothering with end points.

Corollary 14.3.11 Let \( f : (a, b) \to \mathbb{R} \) be continuous and one to one. Suppose \( f' (x_1) \) exists for some \( x_1 \in (a, b) \) and \( f' (x_1) \neq 0 \). Then \( (f^{-1})' (f (x_1)) \) exists and is given by the formula,
\[
(f^{-1})' (f (x_1)) = \frac{1}{f' (x_1)}.
\]
This is one of those theorems which is very easy to remember if you neglect the difficult questions and simply focus on formal manipulations. Consider the following,
\[
f^{-1} (f (x)) = x.
\]
Now use the chain rule on both sides to write
\[(f^{-1})' (f(x)) f'(x) = 1,
\]
and then divide both sides by \(f'(x)\) to obtain
\[(f^{-1})' (f(x)) = \frac{1}{f'(x)}.
\]
Of course this gives the conclusion of the above theorem rather effortlessly and it is formal manipulations like this which aid many of us in remembering formulas such as the one given in the theorem.

### 14.3.2 Independence Of Parameterization

**Theorem 14.3.12** Let \(\phi : [a, b] \to [c, d]\) be one to one and suppose \(\phi '\) exists and is continuous on \([a, b]\). Then if \(f\) is a continuous function defined on \([a, b]\) which is Riemann integrable,\(^2\),
\[
\int_c^d f(s) \, ds = \int_a^b f(\phi(t)) |\phi'(t)| \, dt
\]

**Proof**: Let \(F'(s) = f(s)\). (For example, let \(F(s) = \int_a^s f(r) \, dr\)) Then the first integral equals \(F(d) - F(c)\) by the fundamental theorem of calculus. By Lemma 14.3.8, \(\phi\) is either strictly increasing or strictly decreasing. Suppose \(\phi\) is strictly decreasing. Then \(\phi(a) = d\) and \(\phi(b) = c\). Therefore, \(\phi' \leq 0\) and the second integral equals
\[
\int_a^b f(\phi(t)) \phi'(t) \, dt + \int_b^a \frac{d}{dt} (F(\phi(t))) \, dt = F(\phi(a)) - F(\phi(b)) = F(d) - F(c).
\]
The case when \(\phi\) is increasing is similar. This proves the theorem.

**Lemma 14.3.13** Let \(f : [a, b] \to C\), \(g : [c, d] \to C\) be parameterizations of a smooth curve which satisfy conditions 1 - 5. Then \(\phi(t) \equiv g^{-1} \circ f(t)\) is 1 - 1 on \((a, b)\), continuous on \([a, b]\), and either strictly increasing or strictly decreasing on \([a, b]\).

**Proof**: It is obvious \(\phi\) is 1 - 1 on \((a, b)\) from the conditions \(f\) and \(g\) satisfy. It only remains to verify continuity on \([a, b]\) because then the final claim follows from Lemma 14.3.8. If \(\phi\) is not continuous on \([a, b]\), then there exists a sequence, \(\{t_n\} \subseteq [a, b]\) such that \(t_n \to t\) but \(\phi(t_n)\) fails to converge to \(\phi(t)\). Therefore, for some \(\varepsilon > 0\) there exists a subsequence, still denoted by \(n\) such that \(|\phi(t_n) - \phi(t)| \geq \varepsilon\). Using the sequential compactness of \([c, d]\), (See Theorem 14.3.7 on Page 271.) there is a further subsequence, still denoted by \(n\) such that \(\{\phi(t_n)\}\) converges to a point, \(s\), of \([c, d]\) which is not equal to \(\phi(t)\). Thus \(g^{-1} \circ f(t_n) \to s\) and still \(t_n \to t\). Therefore, the continuity of \(f\) and \(g\) imply \(f(t_n) \to g(s)\) and \(f(t_n) \to f(t)\). Therefore, \(g(s) = f(t)\) and so \(s = g^{-1} \circ f(t) = \phi(t)\), a contradiction. Therefore, \(\phi\) is continuous as claimed.

**Theorem 14.3.14** The length of a smooth curve is not dependent on parameterization.

\(^2\)Recall that all continuous functions of this sort are Riemann integrable.
14.3. INDEPENDENCE OF PARAMETERIZATION

Proof: Let $C$ be the curve and suppose $f : [a,b] \to C$ and $g : [c,d] \to C$ both satisfy conditions 1 - 5. Is it true that $\int_a^b |f'(t)| \, dt = \int_c^d |g'(s)| \, ds$?

Let $\phi(t) \equiv g^{-1} \circ f(t)$ for $t \in [a,b]$. Then by the above lemma $\phi$ is either strictly increasing or strictly decreasing on $[a,b]$. Suppose for the sake of simplicity that it is strictly increasing. The decreasing case is handled similarly.

Let $s_0 \in \phi([a+\delta,b-\delta]) \subset (c,d)$. Then by assumption 4, $g_i'(s_0) \neq 0$ for some $i$. By continuity of $g_i'$, it follows $g_i'(s) \neq 0$ for all $s \in I$ where $I$ is an open interval contained in $[c,d]$ which contains $s_0$. It follows that on this interval, $g_i$ is either strictly increasing or strictly decreasing. Therefore, $J \equiv g_i(I)$ is also an open interval and you can define a differentiable function, $h_i : J \to I$ by

$$h_i(g_i(s)) = s.$$ 

This implies that for $s \in I$,

$$h_i'(g_i(s)) = \frac{1}{g_i'(s)}. \quad (14.3)$$

Now letting $s = \phi(t)$ for $s \in I$, it follows $t \in J_1$, an open interval. Also, for $s$ and $t$ related this way, $f(t) = g(s)$ and so in particular, for $s \in I$,

$$g_i(s) = f_i(t).$$

Consequently,

$$s = h_i(f_i(t)) = \phi(t)$$

and so, for $t \in J_1$,

$$\phi'(t) = h_i'(f_i(t)) f_i'(t) = h_i'(g_i(s)) f_i'(t) = \frac{f_i'(t)}{g_i'(\phi(t))} \quad (14.4)$$

which shows that $\phi'$ exists and is continuous on $J_1$, an open interval containing $\phi^{-1}(s_0)$.

Since $s_0$ is arbitrary, this shows $\phi'$ exists on $[a+\delta,b-\delta]$ and is continuous there.

Now $f(t) = g \circ (g^{-1} \circ f)(t) = g(\phi(t))$ and it was just shown that $\phi'$ is a continuous function on $[a-\delta,b+\delta]$. It follows

$$f'(t) = g'(\phi(t)) \phi'(t)$$

and so, by Theorem 14.3.12,

$$\int_{\phi(a+\delta)}^{\phi(b-\delta)} |g'(s)| \, ds = \int_{a+\delta}^{b-\delta} |g'(\phi(t))| |\phi'(t)| \, dt = \int_{a+\delta}^{b-\delta} |f'(t)| \, dt.$$ 

Now using the continuity of $\phi, g'$, and $f'$ on $[a,b]$ and letting $\delta \to 0+$ in the above, yields

$$\int_a^b |g'(s)| \, ds = \int_a^b |f'(t)| \, dt$$

and this proves the theorem.
Some Curvilinear Coordinate Systems

15.0.3 Outcomes

1. Recall and use polar, cylindrical, and spherical coordinates.

2. Recall and understand the derivation of Kepler’s laws.

3. Recall and apply the concept of acceleration in polar coordinates.

15.1 Polar Cylindrical And Spherical Coordinates

So far points have been identified in terms of Cartesian coordinates but there are other ways of specifying points in two and three dimensional space. These other ways involve using a list of two or three numbers which have a totally different meaning than Cartesian coordinates to specify a point in two or three dimensional space. In general these lists of numbers which have a different meaning than Cartesian coordinates are called Curvilinear coordinates. Probably the simplest curvilinear coordinate system is that of polar coordinates. The idea is suggested in the following picture.

You see in this picture, the number $r$ identifies the distance of the point from the origin, $(0,0)$ while $\theta$ is the angle shown between the positive $x$ axis and the line from the origin to the point. This angle will always be given in radians and is in the interval $[0,2\pi)$. Thus the given point, indicated by a small dot in the picture, can be described in terms of the Cartesian coordinates, $(x,y)$ or the polar coordinates, $(r,\theta)$. How are the two coordinates
systems related? From the picture,

\[ x = r \cos(\theta), \quad y = r \sin(\theta). \quad \text{(15.1)} \]

**Example 15.1.1** The polar coordinates of a point in the plane are \((5, \frac{\pi}{6})\). Find the Cartesian or rectangular coordinates of this point.

From (15.1), \(x = 5 \cos\left(\frac{\pi}{6}\right) = \frac{5}{2}\sqrt{3}\) and \(y = 5 \sin\left(\frac{\pi}{6}\right) = \frac{5}{2}\). Thus the Cartesian coordinates are \((\frac{5}{2}\sqrt{3}, \frac{5}{2})\).

**Example 15.1.2** Suppose the Cartesian coordinates of a point are \((3, 4)\). Find the polar coordinates.

Recall that \(r\) is the distance from \((0, 0)\) and so \(r = 5 = \sqrt{3^2 + 4^2}\). It remains to identify the angle. Note the point is in the first quadrant, (Both the \(x\) and \(y\) values are positive.) Therefore, the angle is something between 0 and \(\pi/2\) and also \(3 = 5 \cos(\theta)\), and \(4 = 5 \sin(\theta)\). Therefore, dividing yields \(\tan(\theta) = 4/3\). At this point, use a calculator or a table of trigonometric functions to find that at least approximately, \(\theta = 0.927 295\) radians.

Now consider two three dimensional generalizations of polar coordinates. The following picture serves as motivation for the definition of these two other coordinate systems.

\[
\begin{array}{c}
\text{(\(x, y_1, z_1\))} \\
\text{(\(r, \theta, z_1\))} \\
\text{\((\rho, \phi, \theta)\)}
\end{array}
\]

In this picture, \(\rho\) is the distance between the origin, the point whose Cartesian coordinates are \((0, 0, 0)\) and the point indicated by a dot and labeled as \((x_1, y_1, z_1)\), \((r, \theta, z_1)\), and \((\rho, \phi, \theta)\). The angle between the positive \(z\) axis and the line between the origin and the point indicated by a dot is denoted by \(\phi\), and \(\theta\), is the angle between the positive \(x\) axis and the line joining the origin to the point \((x_1, y_1, 0)\) as shown, while \(r\) is the length of this line. Thus \(r\) and \(\theta\) determine a point in the plane determined by letting \(z = 0\) and \(r\) and \(\theta\) are the usual polar coordinates. Thus \(r \geq 0\) and \(\theta \in [0, 2\pi]\). Letting \(z_1\) denote the usual \(z\) coordinate of a point in three dimensions, like the one shown as a dot, \((r, \theta, z_1)\) are the cylindrical coordinates of the dotted point. The spherical coordinates are determined by \((\rho, \phi, \theta)\). When \(\rho\) is specified, this indicates that the point of interest is on some sphere of radius \(\rho\) which is centered at the origin. Then when \(\phi\) is given, the location of the point is narrowed down to a circle and finally, \(\theta\) determines which point is on this circle. Let \(\phi \in [0, \pi], \theta \in [0, 2\pi], \) and \(\rho \in [0, \infty)\). The picture shows how to relate these new coordinate
systems to Cartesian coordinates. For Cylindrical coordinates,

\[ x = r \cos(\theta), \]
\[ y = r \sin(\theta), \]
\[ z = z \]

and for spherical coordinates,

\[ x = \rho \sin(\phi) \cos(\theta), \]
\[ y = \rho \sin(\phi) \sin(\theta), \]
\[ z = \rho \cos(\phi). \]

Spherical coordinates should be especially interesting to you because you live on the surface of a sphere. This has been known for several hundred years. You may also know that the standard way to determine position on the earth is to give the longitude and latitude. The latitude corresponds to \( \phi \) and the longitude corresponds to \( \theta \).¹

**Example 15.1.3** Express the surface, \( z = \frac{1}{\sqrt{3}} \sqrt{x^2 + y^2} \) in spherical coordinates.

This is

\[ \rho \cos(\phi) = \frac{1}{\sqrt{3}} \sqrt{(\rho \sin(\phi) \cos(\theta))^2 + (\rho \sin(\phi) \sin(\theta))^2} = \frac{1}{3} \sqrt{3} \rho \sin(\phi). \]

Therefore, this reduces to

\[ \tan(\phi) = \sqrt{3} \]

and so this is just \( \phi = \pi/3 \).

**Example 15.1.4** Express the surface, \( y = x \) in terms of spherical coordinates.

This says \( \rho \sin(\phi) \sin(\theta) = \rho \sin(\phi) \cos(\theta) \). Thus \( \sin(\theta) = \cos(\theta) \). You could also write \( \tan(\theta) = 1 \).

**Example 15.1.5** Express the surface, \( x^2 + y^2 = 4 \) in cylindrical coordinates.

This says \( r^2 \cos^2(\theta) + r^2 \sin^2(\theta) = 4 \). Thus \( r = 2 \).

15.2 The Acceleration In Polar Coordinates

Sometimes you have information about forces which act not in the direction of the coordinate axes but in some other direction. When this is the case, it is often useful to express things in terms of different coordinates which are consistent with these directions. A good example of this is the force exerted by the sun on a planet. This force is always directed toward the sun and so the force vector changes as the planet moves. To discuss this, consider the following simple diagram in which two unit vectors, \( e_r \) and \( e_\theta \) are shown.

¹Actually latitude is determined on maps and in navigation by measuring the angle from the equator rather than the pole but it is essentially the same idea that we have presented here.
The vector, \( \mathbf{e}_r = (\cos \theta, \sin \theta) \) and the vector, \( \mathbf{e}_\theta = (-\sin \theta, \cos \theta) \). You should convince
yourself that the picture above corresponds to this definition of the two vectors. Note that
\( \mathbf{e}_r \) is a unit vector pointing away from 0 and
\[
\mathbf{e}_\theta = \frac{d \mathbf{e}_r}{d \theta}, \quad \mathbf{e}_r = -\frac{d \mathbf{e}_\theta}{d \theta}.
\] (15.2)

Now consider the position vector from 0 of a point in the plane, \( \mathbf{r}(t). \) Then
\[
\mathbf{r}(t) = r(t) \mathbf{e}_r(\theta(t))
\]
where \( r(t) = |\mathbf{r}(t)|. \) Thus \( r(t) \) is just the distance from the origin, 0 to the point. What is
the velocity and acceleration? Using the chain rule,
\[
\frac{d \mathbf{e}_r}{dt} = \frac{d \mathbf{e}_r}{d \theta}(t), \quad \frac{d \mathbf{e}_\theta}{dt} = \frac{d \mathbf{e}_\theta}{d \theta}(t)
\]
and so from (15.2),
\[
\frac{d \mathbf{e}_r}{dt} = \theta'(t) \mathbf{e}_\theta, \quad \frac{d \mathbf{e}_\theta}{dt} = -\theta'(t) \mathbf{e}_r
\] (15.3)

Using (15.3) as needed along with the product rule and the chain rule,
\[
\begin{align*}
\mathbf{r}'(t) &= \mathbf{r}'(t) \mathbf{e}_r + r(t) \frac{d}{dt} \left( \mathbf{e}_r(\theta(t)) \right) \\
&= \mathbf{r}'(t) \mathbf{e}_r + r(t) \theta'(t) \mathbf{e}_\theta.
\end{align*}
\]

Next consider the acceleration.
\[
\begin{align*}
\mathbf{r}''(t) &= r''(t) \mathbf{e}_r + r'(t) \frac{d \mathbf{e}_r}{dt} + r'(t) \theta'(t) \mathbf{e}_\theta + r(t) \theta''(t) \mathbf{e}_\theta + r(t) \theta'(t) \frac{d}{dt} \left( \mathbf{e}_\theta \right) \\
&= r''(t) \mathbf{e}_r + 2r'(t) \theta'(t) \mathbf{e}_\theta + r(t) \theta''(t) \mathbf{e}_\theta + r(t) \theta'(t) \left( -\mathbf{e}_r \right) \theta'(t) \\
&= \left( r''(t) - r(t) \theta'(t)^2 \right) \mathbf{e}_r + \left( 2r'(t) \theta'(t) + r(t) \theta''(t) \right) \mathbf{e}_\theta.
\end{align*}
\] (15.4)

This is a very profound formula. Consider the following examples.

**Example 15.2.1** Suppose an object of mass \( m \) moves at a uniform speed, \( s, \) around a circle
of radius \( R. \) Find the force acting on the object.

By Newton’s second law, the force acting on the object is \( mr'' \). In this case, \( r(t) = R, \) a constant
and since the speed is constant, \( \theta'' = 0. \) Therefore, the term in (15.4) corresponding
to \( \mathbf{e}_\theta \) equals zero and \( mr'' = -R\theta'(t)^2 \mathbf{e}_r. \) The speed of the object is \( s \) and so it moves \( s/R \)
radians in unit time. Thus \( \theta'(t) = s/R \) and so
\[
\begin{align*}
mr'' &= -mR \frac{s^2}{R^2} \mathbf{e}_r = -m \frac{s^2}{R} \mathbf{e}_r.
\end{align*}
\]
This is the familiar formula for centripetal force from elementary physics, obtained as a very special case of (15.4).

Example 15.2.2 A platform rotates at a constant speed in the counter clockwise direction and an object of mass \( m \) moves from the center of the platform toward the edge at constant speed. What forces act on this object?

Let \( v \) denote the constant speed of the object moving toward the edge of the platform. Then

\[
r'(t) = v, \quad r''(t) = 0, \quad \theta''(t) = 0,
\]

while \( \theta'(t) = \omega \), a positive constant. From (15.4)

\[
m r''(t) = -m v^2 \omega^2 e_r + m 2 v \omega e_\theta.
\]

Thus the object experiences centripetal force from the first term and also a funny force from the second term which is in the direction of rotation of the platform. You can observe this by experiment if you like. Go to a playground and have someone spin one of those merry go rounds while you ride it and move from the center toward the edge. The term \( 2r' \theta' \) is called the Coriolis force.

Suppose at each point of space, \( r \) is associated a force, \( F(r) \) which a given object of mass \( m \) will experience if its position vector is \( r \). This is called a force field. A force field is a central force field if \( F(r) = g(r)e_r \). Thus in a central force field, the force an object experiences will always be directed toward or away from the origin, \( 0 \). The following simple lemma is very interesting because it says that in a central force field, objects must move in a plane.

Lemma 15.2.3 Suppose an object moves in three dimensions in such a way that the only force acting on the object is a central force. Then the motion of the object is in a plane.

Proof: Let \( r(t) \) denote the position vector of the object. Then from the definition of a central force and Newton’s second law,

\[
m r'' = g(r) r.
\]

Therefore, \( m r'' \times r = m (r' \times r)' = g(r) r \times r = 0 \). Therefore, \( (r' \times r) = n \), a constant vector and \( \text{so} r \cdot n = r \cdot (r' \times r) = 0 \) showing that \( n \) is a normal vector to a plane which contains \( r(t) \) for all \( t \). This proves the lemma.

The next example has as a special case one of Kepler’s laws, Kepler’s second law, the equal area law.

Example 15.2.4 An object moves in three dimensions in such a way that the only force acting on the object is a central force. Then the object moves in a plane and the radius vector from the origin to the object sweeps out area at a constant rate.

The above lemma says the object moves in a plane. From the assumption that the force field is a central force field, it follows from (15.4) that

\[
2 r' (t) \theta' (t) + r(t) \theta'' (t) = 0
\]

Multiply both sides of this equation by \( r \). This yields

\[
2 r r' \theta' + r^2 \theta'' = (r^2 \theta')' = 0. \quad (15.5)
\]
Consequently,

\[ r^2 \theta' = c \]  \hspace{1cm} (15.6)

for some constant, \( C \). Now consider the following picture.

In this picture, \( d\theta \) is the indicated angle and the two lines determining this angle are position vectors for the object at point \( t \) and point \( t + dt \). The area of the circular sector, \( dA \), is essentially \( r^2 d\theta \) and so \( dA = \frac{1}{2} r^2 d\theta \). Therefore,

\[ \frac{dA}{dt} = \frac{1}{2} r^2 \frac{d\theta}{dt} = \frac{c}{2}. \]  \hspace{1cm} (15.7)

### 15.3 Planetary Motion

Kepler’s laws of planetary motion state that planets move around the sun along an ellipse, the equal area law described above holds, and there is a formula for the time it takes for the planet to move around the sun. These laws, discovered by Kepler, were shown by Newton to be consequences of his law of gravitation which states that the force acting on a mass, \( m \) by a mass, \( M \) is given by

\[ F = -\frac{GMm}{r^2} \]

where \( r \) is the distance between centers of mass and \( r \) is the position vector from \( M \) to \( m \). Here \( G \) is the gravitation constant. This is called an inverse square law. Gravity acts according to this law and so does electrostatic force. The constant, \( G \), is very small when usual units are used and it has been computed using a very delicate experiment. It is now accepted to be

\[ 6.67 \times 10^{-11} \ \text{Newton meter}^2/\text{kilogram}^2. \]

The experiment involved a light source shining on a mirror attached to a quartz fiber from which was suspended a long rod with two equal masses at the ends which were attracted by two larger masses. The gravitation force between the suspended masses and the two large masses caused the fibre to twist ever so slightly and this twisting was measured by observing the deflection of the light reflected from the mirror on a scale placed some distance from the fibre. The constant was first measured successfully by Lord Cavendish in 1798 and the present accepted value was obtained in 1942. Experiments like these are major accomplishments.

In the following argument, \( M \) is the mass of the sun and \( m \) is the mass of the planet. (It could also be a comet or an asteroid.)

Consider the first of Kepler’s laws, the one which states that planets move along ellipses. From Lemma 15.2.3, the motion is in a plane. Now from (15.4) and Newton’s second law,

\[ \left( r''(t) - r(t) \theta'(t)^2 \right) e_r + \left( 2r'(t) \theta'(t) + r(t) \theta''(t) \right) e_\theta = -\frac{GMm}{m} \left( \frac{1}{r^2} \right) e_r = -k \left( \frac{1}{r^2} \right) e_r \]
15.3. PLANETARY MOTION

Thus \( k = GM \) and

\[
r''(t) - r(t) \theta'(t)^2 = -k \left( \frac{1}{r^2} \right), \quad 2r'(t) \theta'(t) + r(t) \theta''(t) = 0. \tag{15.8}
\]

As in (15.5), \((r^2 \theta')' = 0\) and so there exists a constant, \(c\), such that

\[
r^2 \theta' = c. \tag{15.9}
\]

Therefore, also,

\[
2rr' \theta' + r^2 \theta'' = 0
\]

and so

\[
\theta'' = \frac{-2rr' \theta'}{r^2} = \frac{-2r' \theta'}{r} = -\frac{2}{r} \frac{dr}{dt} c \frac{r^2}{dt}
\]

(15.10)

(15.11)

Now consider the first of the above equations. The question of interest is to know how \(r\) depends on \(\theta\). By the chain rule, regarding \(r\) as a function of \(\theta\) and \(\theta\) as a function of \(t\),

\[
\frac{dr}{d\theta} \frac{d\theta}{dt} = \frac{dr}{dt}
\]

(15.12)

and by (15.9),

\[
\frac{dr}{d\theta} = \frac{c}{r^2} \frac{dr}{dt}
\]

(15.13)

Also, by (15.10) and (15.9),

\[
\frac{d^2 \theta}{dt^2} = \frac{-2c}{r^3} \frac{dr}{dt} = \frac{-2c}{r^3} \left( \frac{dr}{d\theta} \frac{d\theta}{dt} \right)
\]

\[
= \frac{-2c}{r^3} \frac{dr}{d\theta} \left( \frac{c}{r^2} \right) = -\frac{2c^2}{r^5} \frac{dr}{d\theta}
\]

Differentiating (15.12) again with respect to \(t\),

\[
\frac{d^2 r}{dt^2} = \frac{d^2 r}{d\theta^2} \left( \frac{d\theta}{dt} \right)^2 + \frac{dr}{d\theta} \frac{d^2 \theta}{dt^2}
\]

\[
= \frac{d^2 r}{d\theta^2} \left( \frac{c}{r^2} \right)^2 + \frac{dr}{d\theta} \left( -\frac{2c^2}{r^5} \frac{dr}{d\theta} \right)
\]

\[
= \frac{d^2 r}{d\theta^2} \left( \frac{c}{r^2} \right)^2 - \left( \frac{dr}{d\theta} \right)^2 \left( \frac{2c^2}{r^5} \right).
\]

It follows that the first equation of (15.8) yields

\[
\frac{d^2 r}{d\theta^2} \left( \frac{c}{r^2} \right)^2 - \left( \frac{dr}{d\theta} \right)^2 \left( \frac{2c^2}{r^5} \right) - r \left( \frac{c}{r^2} \right)^2 = -k \left( \frac{1}{r^2} \right)
\]

which is a fairly messy looking differential equation. However, it can be simplified by multiplying both sides by \(\frac{1}{2c^2} r^2\) to get

\[
\frac{d^2 r}{d\theta^2} - \left( \frac{dr}{d\theta} \right)^2 \left( \frac{2}{r^2} \right) - r = -\frac{k}{c^2} r^2
\]

(15.14)
Next consider the above equation in terms of $\rho = r^{-1}$. Thus, from the chain rule,

$$r = \rho^{-1}, \quad \frac{dr}{d\theta} = (-1) \rho^{-2} \frac{d\rho}{d\theta},$$

$$\frac{d^2r}{d\theta^2} = 2\rho^{-3} \left( \frac{d\rho}{d\theta} \right)^2 - \rho^{-2} \frac{d^2\rho}{d\theta^2}.$$ 

substituting this into (15.14),

$$\frac{d^2\rho}{d\theta^2} \left( \frac{d\rho}{d\theta} \right)^2 - \rho^{-2} \frac{d^2\rho}{d\theta^2} = 2\left( \frac{d\rho}{d\theta} \right)^2 - \rho^{-2} \frac{d^2\rho}{d\theta^2} = \left( \frac{1}{2} \right) \left( \frac{d\rho}{d\theta} \right)^2 (2\rho) - \rho^{-1} = -\frac{k}{\rho^2 c^2}.$$ 

Now note that the first and third terms add to zero and so

$$-\rho^{-2} \frac{d^2\rho}{d\theta^2} - \rho^{-1} = -\frac{k}{\rho^2 c^2}.$$ 

Multiplying both sides by $-\rho^{-2}$ yields the equation,

$$\frac{d^2\rho}{d\theta^2} + \rho = \frac{k}{c^2},$$

a much more manageable equation. Multiply both sides by $\frac{d\rho}{d\theta}$.

$$\frac{d^2\rho}{d\theta^2} + \rho \frac{d\rho}{d\theta} = \frac{k}{c^2} \frac{d\rho}{d\theta}.$$ 

Then from the product rule,

$$\frac{1}{2} \frac{d}{d\theta} \left( \left( \frac{d\rho}{d\theta} \right)^2 + \rho^2 \right) = \frac{k}{c^2} \frac{d\rho}{d\theta}.$$ 

Therefore, there exists a constant, $c_1$ such that

$$\frac{1}{2} \left( \left( \frac{d\rho}{d\theta} \right)^2 + \rho^2 \right) - \frac{k}{c^2} \rho = c_1$$

and so

$$\left( \frac{d\rho}{d\theta} \right)^2 = 2c_1 + \frac{k}{c^2} \rho - \rho^2$$

$$= \left( \frac{k^2}{4c^4} + 2c_1 \right) - \left( \rho - \frac{k}{2c^2} \right)^2$$

$$\equiv \delta^2 - \left( \rho - \frac{k}{2c^2} \right)^2$$

Now letting $\rho_1 = \rho - \frac{k}{2c^2},$

$$\frac{1}{\delta^2} \left( \frac{d\rho_1}{d\theta} \right)^2 + \left( \frac{\rho_1}{\delta} \right)^2 = 1.$$
which shows that \( \left( \frac{1}{\delta} \frac{d\rho_1}{d\theta}, \frac{\rho_1}{\delta} \right) \) is a point on the unit circle. Therefore, there exists an angle, \( \alpha (\theta) \) such that
\[
\frac{d\rho_1}{d\theta} = \delta \cos (\alpha (\theta)), \rho_1 = \delta \sin (\alpha (\theta)).
\]
Differentiating the second equation with respect to \( \theta \),
\[
\frac{d\rho_1}{d\theta} = \alpha' (\theta) \delta \cos (\alpha (\theta)), \rho_1 = \delta \sin (\alpha (\theta)).
\]
and so \( \alpha' (\theta) = 1 \). Therefore, \( \alpha (\theta) = \theta + \phi \) for some constant, \( \phi \). Redefining, \( \theta \) if necessary, (Let \( \tilde{\theta} = \theta + \phi \) it can be assumed that \( \phi = 0 \) so
\[
\rho - \frac{k}{2c^2} = \rho_1 = \delta \sin \theta.
\]
Thus
\[
\rho = \frac{k}{c^2} + \delta \sin \theta
\]
and so
\[
r = \frac{1}{\sqrt{\frac{1}{\delta^2} + \delta \sin \theta}} = \frac{c^2/k}{1 + (c^2/k) \delta \sin \theta}
\]
\[
= \frac{pe}{1 + \varepsilon \sin \theta}
\]
where
\[
\varepsilon = (c^2/k) \delta \text{ and } p = c^2/k \varepsilon.
\]
Here all these constants are nonnegative.

Thus
\[
r + \varepsilon r \sin \theta = ep
\]
and so \( r = (ep - \varepsilon y) \). Then squaring both sides,
\[
x^2 + y^2 = (ep - \varepsilon y)^2 = \varepsilon^2 p^2 - 2pe^2 y + \varepsilon^2 y^2
\]
And so
\[
x^2 + (1 - \varepsilon^2) y^2 = \varepsilon^2 p^2 - 2pe^2 y.
\]
(15.17)
In case \( \varepsilon = 1 \), this reduces to the equation of a parabola. If \( \varepsilon < 1 \), this reduces to the equation of an ellipse and if \( \varepsilon > 1 \), this is called a hyperbola. This proves that objects which are acted on only by a force of the form given in the above example move along hyperbolas, ellipses or circles. The case where \( \varepsilon = 0 \) corresponds to a circle. The constant, \( \varepsilon \) is called the eccentricity. This is called Kepler’s first law in the case of a planet.

Kepler’s third law involves the time it takes for the planet to orbit the sun. From (15.17) you can complete the square and obtain
\[
x^2 + (1 - \varepsilon^2) \left( y + \frac{pe^2}{1 - \varepsilon^2} \right)^2 = \varepsilon^2 p^2 + \frac{p^2 \varepsilon^4}{(1 - \varepsilon^2)} = \frac{\varepsilon^2 p^2}{(1 - \varepsilon^2)},
\]
and this yields
\[
x^2 / \left( \frac{\varepsilon^2 p^2}{1 - \varepsilon^2} \right) + \left( y + \frac{pe^2}{1 - \varepsilon^2} \right)^2 / \left( \frac{\varepsilon^2 p^2}{(1 - \varepsilon^2)} \right) = 1.
\]
(15.18)
286

**SOME CURVILINEAR COORDINATE SYSTEMS**

Now note this is the equation of an ellipse and that the diameter of this ellipse is

\[
\frac{2\varepsilon p}{(1 - \varepsilon^2)} \equiv 2a. \tag{15.19}
\]

This follows because

\[
\frac{\varepsilon^2 p^2}{(1 - \varepsilon^2)^2} \geq \frac{\varepsilon^2 p^2}{1 - \varepsilon^2}.
\]

Now let \( T \) denote the time it takes for the planet to make one revolution about the sun. Using this formula, and (15.7) the following equation must hold.

\[
\text{area of ellipse} \left( \pi \frac{\varepsilon p}{\sqrt{1 - \varepsilon^2}} \right) = T \frac{c}{2}.
\]

Therefore,

\[
T = 2 \frac{\pi \varepsilon^2 p^2}{c (1 - \varepsilon^2)^{3/2}}
\]

and so

\[
T^2 = \frac{4\pi^2 \varepsilon^4 p^4}{c^2 (1 - \varepsilon^2)^3}
\]

Now using (15.16),

\[
T^2 = \frac{4\pi^2 \varepsilon^4 p^4}{k \varepsilon p (1 - \varepsilon^2)^3} = \frac{4\pi^2 (\varepsilon p)^3}{k (1 - \varepsilon^2)^3} = \frac{4\pi^2 a^3}{k} = \frac{4\pi^2 a^3}{GM}.
\]

Written more memorably, this has shown

\[
T^2 = \frac{4\pi^2}{GM} \left( \frac{\text{diameter of ellipse}}{2} \right)^3. \tag{15.19}
\]

This relationship is known as Kepler’s third law. Kepler’s second law, the equal area formula, holds for any central force, not just one which satisfies an inverse square law.

### 15.4 Exercises

1. In general it is a stupid idea to try to use algebra to invert and solve for a set of curvilinear coordinates such as polar or cylindrical coordinates in term of Cartesian coordinates. Not only is it often very difficult or even impossible to do it, but also it takes you in entirely the wrong direction because the whole point of introducing the new coordinates is to write everything in terms of these new coordinates and not in terms of Cartesian coordinates. However, sometimes this inversion can be done. Describe how to solve for \( r \) and \( \theta \) in terms of \( x \) and \( y \) in polar coordinates.

2. A point has Cartesian coordinates, \((1, 2, 3)\). Find its spherical and cylindrical coordinates.

3. Describe the following surface in rectangular coordinates. \( \phi = \pi/4 \) where \( \phi \) is the polar angle in spherical coordinates.
4. Describe the following surface in rectangular coordinates. \( \theta = \pi/4 \) where \( \theta \) is the angle measured from the positive \( x \) axis spherical coordinates.

5. Describe the following surface in rectangular coordinates. \( \theta = \pi/4 \) where \( \theta \) is the angle measured from the positive \( x \) axis cylindrical coordinates.

6. Describe the following surface in rectangular coordinates. \( r = 5 \) where \( r \) is one of the cylindrical coordinates.

7. Describe the following surface in rectangular coordinates. \( \rho = 4 \) where \( \rho \) is the distance to the origin.

8. Give the cone, \( z = \sqrt{x^2 + y^2} \) in cylindrical coordinates and in spherical coordinates.

9. Write the following in spherical coordinates.
   (a) \( z = x^2 + y^2 \).
   (b) \( x^2 - y^2 = 1 \).
   (c) \( z^2 + x^2 + y^2 = 6 \).
   (d) \( z = \sqrt{x^2 + y^2} \).
   (e) \( y = x \).
   (f) \( z = x \).

10. Write the following in cylindrical coordinates.
    (a) \( z = x^2 + y^2 \).
    (b) \( x^2 - y^2 = 1 \).
    (c) \( z^2 + x^2 + y^2 = 6 \).
    (d) \( z = \sqrt{x^2 + y^2} \).
    (e) \( y = x \).
    (f) \( z = x \).

11. Suppose you know how the spherical coordinates of a moving point change as a function of \( t \). Can you figure out the velocity of the point? Specifically, suppose \( \phi (t) = t \), \( \theta (t) = 1 + t \), and \( \rho (t) = t \). Find the speed and the velocity of the object in terms of Cartesian coordinates. \textbf{Hint:} You would need to find \( x' (t) \), \( y' (t) \), and \( z' (t) \). Then in terms of Cartesian coordinates, the velocity would be \( x' (t) \mathbf{i} + y' (t) \mathbf{j} + z' (t) \mathbf{k} \).

12. Explain why low pressure areas rotate counter clockwise in the Northern hemisphere and clockwise in the Southern hemisphere. \textbf{Hint:} Note that from the point of view of an observer fixed in space, the low pressure area already has a counter clockwise rotation because of the rotation of the earth and its spherical shape. Now consider (15.6). In the low pressure area stuff will move toward the center so \( r \) gets smaller. How are things different in the Southern hemisphere?

13. What are some physical assumptions which are made in the above derivation of Keplers laws from Newton’s laws of motion?

14. The orbit of the earth is pretty nearly circular and the distance from the sun to the earth is about \( 149 \times 10^6 \) kilometers. Using (15.19) and the above value of the universal gravitation constant, determine the mass of the sun. The earth goes around it in 365 days. (Actually it is 365.256 days.)
15. It is desired to place a satellite above the equator of the earth which will rotate about the center of mass of the earth every 24 hours. Is it necessary that the orbit be circular? What if you want the satellite to stay above the same point on the earth at all times? If the orbit is to be circular and the satellite is to stay above the same point, at what distance from the center of mass of the earth should the satellite be? You may use that the mass of the earth is $5.98 \times 10^{24}$ kilograms. Such a satellite is called geosynchronous.
Functions Of Many Variables

16.0.1 Outcomes

1. Represent a function of two variables by level curves.
2. Identify the characteristics of a function from a graph of its level curves.
3. Recall and use the concept of limit point.
4. Describe the geometrical significance of a directional derivative.
5. Give the relationship between partial derivatives and directional derivatives.
6. Compute partial derivatives and directional derivatives from their definitions.
7. Evaluate higher order partial derivatives.
8. State conditions under which mixed partial derivatives are equal.
10. Describe the gradient of a scalar valued function and use to compute the directional derivative.
11. Explain why the directional derivative is maximized in the direction of the gradient and minimized in the direction of minus the gradient.

16.1 The Graph Of A Function Of Two Variables

With vector valued functions of many variables, it doesn’t take long before it is impossible to draw meaningful pictures. This is because one needs more than three dimensions to accomplish the task and we can only visualize things in three dimensions. Ultimately, one of the main purposes of calculus is to free us from the tyranny of art. In calculus, we are permitted and even required to think in a meaningful way about things which cannot be drawn. However, it is certainly interesting to consider some things which can be visualized and this will help to formulate and understand more general notions which make sense in contexts which cannot be visualized. One of these is the concept of a scalar valued function of two variables.

Let \( f (x, y) \) denote a scalar valued function of two variables evaluated at the point \((x, y)\). Its graph consists of the set of points, \((x, y, z)\) such that \( z = f (x, y) \). How does one go about depicting such a graph? The usual way is to fix one of the variables, say \( x \) and consider the function \( z = f (x, y) \) where \( y \) is allowed to vary and \( x \) is fixed. Graphing this would give a curve which lies in the surface to be depicted. Then do the same thing for other
values of \( x \) and the result would depict the graph desired graph. Computers do this very well. The following is the graph of the function \( z = \cos(x) \sin(2x + y) \) drawn using Maple, a computer algebra system.\(^1\).

Notice how elaborate this picture is. The lines in the drawing correspond to taking one of the variables constant and graphing the curve which results. The computer did this drawing in seconds but you couldn't do it as well if you spent all day on it. I used a grid consisting of 70 choices for \( x \) and 70 choices for \( y \).

Sometimes attempts are made to understand three dimensional objects like the above graph by looking at contour graphs in two dimensions. The contour graph of the above three dimensional graph is below and comes from using the computer algebra system again.

This is in two dimensions and the different lines in two dimensions correspond to points on the three dimensional graph which have the same \( z \) value. If you have looked at a weather map, these lines are called isotherms or isobars depending on whether the function involved is temperature or pressure. In a contour geographic map, the contour lines represent constant altitude. If many contour lines are close to each other, this indicates rapid change in the altitude, temperature, pressure, or whatever else may be measured.

A scalar function of three variables, cannot be visualized because four dimensions are required. However, some people like to try and visualize even these examples. This is done by looking at level surfaces in \( \mathbb{R}^3 \) which are defined as surfaces where the function assumes a constant value. They play the role of contour lines for a function of two variables. As a simple example, consider \( f(x, y, z) = x^2 + y^2 + z^2 \). The level surfaces of this function would be concentric spheres centered at \( 0 \). (Why?) Another way to visualize objects in higher dimensions involves the use of color and animation. However, there really are limits to what you can accomplish in this direction. So much for art.

However, the concept of level curves is quite useful because these can be drawn.

**Example 16.1.1** Determine from a contour map where the function, \( f(x, y) = \sin(x^2 + y^2) \) is steepest.

\(^1\)I used Maple and exported the graph as an eps. file which I then imported into this document.
In the picture, the steepest places are where the contour lines are close together because they correspond to various values of the function. You can look at the picture and see where they are close and where they are far. This is the advantage of a contour map.

16.2 Review Of Limits

Recall the concept of limit of a function of many variables. When \( f : D(f) \to \mathbb{R}^q \) one can only consider in a meaningful way limits at limit points of the set, \( D(f) \).

**Definition 16.2.1** Let \( A \) denote a nonempty subset of \( \mathbb{R}^p \). A point, \( x \) is said to be a limit point of the set, \( A \) if for every \( r > 0 \), \( B(x, r) \) contains infinitely many points of \( A \).

**Example 16.2.2** Let \( S \) denote the set, \( \{(x, y, z) \in \mathbb{R}^3 : x, y, z \text{ are all in } \mathbb{N}\} \). Which points are limit points?

This set does not have any because any two of these points are at least as far apart as 1. Therefore, if \( x \) is any point of \( \mathbb{R}^3 \), \( B(x, 1/4) \) contains at most one point.

**Example 16.2.3** Let \( U \) be an open set in \( \mathbb{R}^3 \). Which points of \( U \) are limit points of \( U \)?

They all are. From the definition of \( U \) being open, if \( x \in U \), there exists \( B(x, r) \subseteq U \) for some \( r > 0 \). Now consider the line segment \( x + t\mathbf{e}_1 \) where \( t \in [0, 1/2] \). This describes infinitely many points and they are all in \( B(x, r) \) because

\[
|x + t\mathbf{e}_1 - x| = tr < r.
\]

Therefore, every point of \( U \) is a limit point of \( U \).

The case where \( U \) is open will be the one of most interest but many other sets have limit points.

**Definition 16.2.4** Let \( f : D(f) \subseteq \mathbb{R}^p \to \mathbb{R}^q \) where \( q, p \geq 1 \) 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,

\[
|L - f(y)| < \varepsilon.
\]
The condition that \( x \) must be a limit point of \( D(f) \) if you are to take a limit at \( x \) is what makes the limit well defined.

**Proposition 16.2.5** Let \( f : D(f) \subseteq \mathbb{R}^p \to \mathbb{R}^q \) where \( q, p \geq 1 \) be a function and let \( x \) be a limit point of \( D(f) \). Then if \( \lim_{y \to x} f(y) \) exists, it must be unique.

**Proof:** Suppose \( \lim_{y \to x} f(y) = L_1 \) and \( \lim_{y \to x} f(y) = L_2 \). Then for \( \varepsilon > 0 \) given, let \( \delta_i > 0 \) correspond to \( L_i \) in the definition of the limit and let \( \delta = \min(\delta_1, \delta_2) \). Since \( x \) is a limit point, there exists \( y \in B(x, \delta) \cap D(f) \). Therefore,

\[
|L_1 - L_2| \leq |L_1 - f(y)| + |f(y) - L_2| < \varepsilon + \varepsilon = 2\varepsilon.
\]

Since \( \varepsilon > 0 \) is arbitrary, this shows \( L_1 = L_2 \). The following theorem summarized many important interactions involving continuity. Most of this theorem has been proved in Theorem 12.4.5 on Page 213 and Theorem 12.4.7 on Page 215.

**Theorem 16.2.6** Suppose \( x \) is a limit point of \( D(f) \) and \( \lim_{y \to x} f(y) = L \), \( \lim_{y \to x} g(y) = K \) where \( K \) and \( L \) are vectors in \( \mathbb{R}^p \) for \( p \geq 1 \). Then if \( a, b \in \mathbb{R} \),

\[
\lim_{y \to x} af(y) + bg(y) = aL + bK, \quad (16.1)
\]

\[
\lim_{y \to x} f \cdot g(y) = L \cdot K \quad (16.2)
\]

Also, if \( h \) is a continuous function defined near \( L \), then

\[
\lim_{y \to x} h \circ f(y) = h(L). \quad (16.3)
\]

For a vector valued function, \( f(y) = (f_1(y), \cdots, f_q(y)) \), \( \lim_{y \to x} f(y) = L = (L_1, \cdots, L_k)^T \) if and only if

\[
\lim_{y \to x} f_k(y) = L_k \quad (16.4)
\]

for each \( k = 1, \cdots, p \).

In the case where \( f \) and \( g \) have values in \( \mathbb{R}^3 \)

\[
\lim_{y \to x} f(y) \times g(y) = L \times K. \quad (16.5)
\]

Also recall Theorem 12.4.6 on Page 215.

**Theorem 16.2.7** For \( f : D(f) \to \mathbb{R}^q \) and \( x \in D(f) \) such that \( x \) is a limit point of \( D(f) \), it follows \( f \) is continuous at \( x \) if and only if \( \lim_{y \to x} f(y) = f(x) \).

### 16.3 The Directional Derivative And Partial Derivatives

#### 16.3.1 The Directional Derivative

The directional derivative is just what its name suggests. It is the derivative of a function in a particular direction. The following picture illustrates the situation in the case of a
16.3. THE DIRECTIONAL DERIVATIVE AND PARTIAL DERIVATIVES

function of two variables.

In this picture, \( \mathbf{v} \equiv (v_1, v_2) \) is a unit vector in the \( xy \) plane and \( x_0 \equiv (x_0, y_0) \) is a point in the \( xy \) plane. When \( (x, y) \) moves in the direction of \( \mathbf{v} \), this results in a change in \( z = f(x, y) \) as shown in the picture. The directional derivative in this direction is defined as

\[
\lim_{t \to 0} \frac{f(x_0 + tv_1, y_0 + tv_2) - f(x_0, y_0)}{t}.
\]

It tells how fast \( z \) is changing in this direction. If you looked at it from the side, you would be getting the slope of the indicated tangent line. A simple example of this is a person climbing a mountain. He could go various directions, some steeper than others. The directional derivative is just a measure of the steepness in a given direction. This motivates the following general definition of the directional derivative.

**Definition 16.3.1** Let \( f : U \to \mathbb{R} \) where \( U \) is an open set in \( \mathbb{R}^n \) and let \( \mathbf{v} \) be a unit vector. For \( x \in U \), define the directional derivative of \( f \) in the direction, \( \mathbf{v} \), at the point \( x \) as

\[
D_{\mathbf{v}} f(x) \equiv \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t}.
\]

**Example 16.3.2** Find the directional derivative of the function, \( f(x, y) = x^2 y \) in the direction of \( \mathbf{i} + \mathbf{j} \) at the point \( (1, 2) \).

First you need a unit vector which has the same direction as the given vector. This unit vector is \( \mathbf{v} \equiv \left( \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right) \). Then to find the directional derivative from the definition, write the difference quotient described above. Thus \( f(x + tv) = \left( 1 + \frac{1}{\sqrt{2}} \right)^2 \left( 2 + \frac{1}{\sqrt{2}} \right) \) and \( f(x) = 2 \). Therefore,

\[
\frac{f(x + tv) - f(x)}{t} = \left( 1 + \frac{1}{\sqrt{2}} \right)^2 \left( 2 + \frac{1}{\sqrt{2}} \right) - 2,
\]

and to find the directional derivative, you take the limit of this as \( t \to 0 \). However, this difference quotient equals \( \frac{1}{4} \sqrt{2} \left( 10 + 4t\sqrt{2} + t^2 \right) \) and so, letting \( t \to 0 \),

\[
D_{\mathbf{v}} f(1, 2) = \left( \frac{5}{2} \sqrt{2} \right).
\]
There is something you must keep in mind about this. The direction vector must always be a unit vector\(^2\).

### 16.3.2 Partial Derivatives

There are some special unit vectors which come to mind immediately. These are the vectors, \( \mathbf{e}_i \) where

\[
\mathbf{e}_i = (0, \cdots, 0, 1, 0, \cdots, 0)^T
\]

and the 1 is in the \( i \)th position.

**Definition 16.3.3** Let \( U \) be an open subset of \( \mathbb{R}^n \) and let \( f : U \rightarrow \mathbb{R} \). Then letting \( \mathbf{x} = (x_1, \cdots, x_n)^T \) be a typical element of \( \mathbb{R}^n \),

\[
\frac{\partial f}{\partial x_i}(\mathbf{x}) \equiv D_{\mathbf{e}_i} f(\mathbf{x}).
\]

This is called the partial derivative of \( f \). Thus,

\[
\frac{\partial f}{\partial x_i}(\mathbf{x}) \equiv \lim_{t \to 0} \frac{f(\mathbf{x}+t\mathbf{e}_i) - f(\mathbf{x})}{t} = \lim_{t \to 0} \frac{f(x_1, \cdots, x_i + t, \cdots, x_n) - f(x_1, \cdots, x_i, \cdots, x_n)}{t},
\]

and to find the partial derivative, differentiate with respect to the variable of interest and regard all the others as constants. Other notation for this partial derivative is \( f_{x_i}, f_{i}, \) or \( D_i f \). If \( y = f(\mathbf{x}) \), the partial derivative of \( f \) with respect to \( x_i \) may also be denoted by

\[
\frac{\partial y}{\partial x_i} \text{ or } y_{x_i}.
\]

**Example 16.3.4** Find \( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \) and \( \frac{\partial f}{\partial x} \) if \( f(x, y) = y \sin x + x^2y + z. \)

From the definition above, \( \frac{\partial f}{\partial x} = y \cos x + 2xy, \frac{\partial f}{\partial y} = \sin x + x^2, \) and \( \frac{\partial f}{\partial z} = 1. \) Having taken one partial derivative, there is no reason to stop doing it. Thus, one could take the partial derivative with respect to \( y \) of the partial derivative with respect to \( x \), denoted by \( \frac{\partial^2 f}{\partial y \partial x} \) or \( f_{xy} \). In the above example,

\[
\frac{\partial^2 f}{\partial y \partial x} = f_{xy} = \cos x + 2x.
\]

Also observe that

\[
\frac{\partial^2 f}{\partial x \partial y} = f_{yx} = \cos x + 2x.
\]

Higher order partial derivatives are defined by analogy to the above. Thus in the above example,

\[
f_{yxx} = -\sin x + 2.
\]

These partial derivatives, \( f_{xy} \) are called mixed partial derivatives.

There is an interesting relationship between the directional derivatives and the partial derivatives, provided the partial derivatives exist and are continuous.

\(^2\)Actually, there is a more general formulation of this known as the Gateaux derivative in which the length of \( v \) is not considered but it will not be considered here.
16.3. The Directional Derivative and Partial Derivatives

**Definition 16.3.5** Suppose \( f : U \subseteq \mathbb{R}^n \to \mathbb{R} \) where \( U \) is an open set and the partial derivatives of \( f \) all exist and are continuous on \( U \). Under these conditions, define the gradient of \( f \) denoted \( \nabla f (x) \) to be the vector

\[
\nabla f (x) = (f_{x_1} (x), f_{x_2} (x), \ldots, f_{x_n} (x))^T.
\]

**Proposition 16.3.6** In the situation of Definition 16.3.5 and for \( \mathbf{v} \) a unit vector,

\[
D_\mathbf{v} f (x) = \nabla f (x) \cdot \mathbf{v}.
\]

This proposition will be proved in a more general setting later. For now, you can use it to compute directional derivatives.

**Example 16.3.7** Find the directional derivative of the function, \( f (x, y) = \sin (2x^2 + y^3) \) at \((1,1)\) in the direction \( \left( \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)^T \).

First find the gradient.

\[
\nabla f (x, y) = (4x \cos (2x^2 + y^3), 3y^2 \cos (2x^2 + y^3))^T.
\]

Therefore,

\[
\nabla f (1,1) = (4 \cos (3), 3 \cos (3))^T
\]

The directional derivative is therefore,

\[
(4 \cos (3), 3 \cos (3))^T \cdot \left( \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)^T = \frac{7}{2} (\cos 3) \sqrt{2}.
\]

Another important observation is that the gradient gives the direction in which the function changes most rapidly.

**Proposition 16.3.8** In the situation of Definition 16.3.5, suppose \( \nabla f (x) \neq 0 \). Then the direction in which \( f \) increases most rapidly, that is the direction in which the directional derivative is largest, is the direction of the gradient. Thus \( \mathbf{v} = \nabla f (x) / |\nabla f (x)| \) is the unit vector which maximizes \( D_\mathbf{v} f (x) \) and this maximum value is \( |\nabla f (x)| \). Similarly, \( \mathbf{v} = -\nabla f (x) / |\nabla f (x)| \) is the unit vector which minimizes \( D_\mathbf{v} f (x) \) and this minimum value is \(-|\nabla f (x)|\).

**Proof:** Let \( \mathbf{v} \) be any unit vector. Then from Proposition 16.3.6,

\[
D_\mathbf{v} f (x) = \nabla f (x) \cdot \mathbf{v} = |\nabla f (x)| |\mathbf{v}| \cos \theta = |\nabla f (x)| \cos \theta
\]

where \( \theta \) is the included angle between these two vectors, \( \nabla f (x) \) and \( \mathbf{v} \). Therefore, \( D_\mathbf{v} f (x) \) is maximized when \( \cos \theta = 1 \) and minimized when \( \cos \theta = -1 \). The first case corresponds to the angle between the two vectors being 0 which requires they point in the same direction in which case, it must be that \( \mathbf{v} = \nabla f (x) / |\nabla f (x)| \) and \( D_\mathbf{v} f (x) = |\nabla f (x)| \). The second case occurs when \( \theta = \pi \) and in this case the two vectors point in opposite directions and the directional derivative equals \(-|\nabla f (x)|\).

The concept of a directional derivative for a vector valued function is also easy to define although the geometric significance expressed in pictures is not.

**Definition 16.3.9** Let \( f : U \to \mathbb{R}^p \) where \( U \) is an open set in \( \mathbb{R}^n \) and let \( \mathbf{v} \) be a unit vector. For \( x \in U \), define the directional derivative of \( f \) in the direction, \( \mathbf{v} \), at the point \( x \) as

\[
D_\mathbf{v} f (x) = \lim_{t \to 0} \frac{f (x + tv) - f (x)}{t}.
\]
Example 16.3.10 Let \( f(x,y) = (xy^2, yz)^T \). Find the directional derivative in the direction \((1,2)^T\) at the point \((x,y)\).

First, a unit vector in this direction is \((1/\sqrt{5}, 2/\sqrt{5})^T\) and from the definition, the desired limit is

\[
\lim_{t \to 0} \left( (x + t (1/\sqrt{5})) (y + t (2/\sqrt{5}))^2 - xy^2, (x + t (1/\sqrt{5}))(y + t (2/\sqrt{5})) - xy \right)
\]

\[
= \lim_{t \to 0} \left( \frac{4}{5} xy \sqrt{5} + \frac{4}{5} xt + \frac{1}{5} \sqrt{5} y^2 + \frac{4}{5} ty + \frac{4}{25} t^2 \sqrt{5}, \frac{2}{5} x \sqrt{5} + \frac{1}{5} y \sqrt{5} + \frac{2}{5} t \right)
\]

You see from this example and the above definition that all you have to do is to form the vector which is obtained by replacing each component of the vector with its directional derivative. In particular, you can take partial derivatives of vector valued functions and use the same notation.

Example 16.3.11 Find the partial derivative with respect to \( x \) of the function \( f(x, y, z, w) = (xy^2, z \sin(xy), z^3 x)^T \).

From the above definition, \( f_x(x,y,z) = D_1 f(x,y,z) = (y^2, z \cos(xy), z^3)^T \).

16.4 Mixed Partial Derivatives

Under certain conditions the mixed partial derivatives will always be equal. This astonishing fact is due to Euler in 1734.

Theorem 16.4.1 Suppose \( f : U \subseteq \mathbb{R}^2 \to \mathbb{R} \) 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 \( |t|, |s| < r/2 \) and consider

\[
\Delta (s,t) \equiv \frac{1}{st} \left( f(x + t, y + s) - f(x + t, y) - (f(x, y + s) - f(x, y)) \right).
\]

Note that \((x + t, y + s) \in U\) because

\[
|(x + t, y + s) - (x,y)| = |(t,s)| = (t^2 + s^2)^{1/2} \leq \left( \frac{r^2}{4} + \frac{r^2}{4} \right)^{1/2} = \frac{r}{\sqrt{2}} < r.
\]

As implied above, \( h(t) \equiv f(x + t, y + s) - f(x + t, y) \). Therefore, by the mean value theorem from calculus and the (one variable) chain rule,

\[
\Delta (s,t) = \frac{1}{st} (h(t) - h(0)) = \frac{1}{s} h'(x + at) t = \frac{1}{s} (f_x(x + at, y + s) - f_x(x + at, y))
\]
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 (16.6), \( \Delta (s, t) \) is also 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 \( (x, y) \),
\[
\lim_{(s,t)\to(0,0)} \Delta (s, t) = f_{xy} (x, y) = f_{yx} (x, y).
\]

This proves the theorem.

The following is obtained from the above by simply fixing all the variables except for the two of interest.

**Corollary 16.4.2** Suppose \( U \) is an open subset of \( \mathbb{R}^n \) and \( f : U \to \mathbb{R} \) has the property that for two indices, \( k, l, f_{x_k}, f_{x_l}, f_{x_kx_l}, \) 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 [3].

**Example 16.4.3** 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 \( (x, y) \neq (0, 0) \),
\[
f_x = y \frac{x^4 - y^4 + 4x^2y^2}{(x^2 + y^2)^2}, \quad f_y = x \frac{x^4 - y^4 - 4x^2y^2}{(x^2 + y^2)^2}
\]

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.
16.5 Partial Differential Equations

Partial differential equations are equations which involve the partial derivatives of some function. The most famous partial differential equations involve the Laplacian, named after Laplace\(^3\).

**Definition 16.5.1** Let \( u \) be a function of \( n \) variables. Then \( \Delta u \equiv \sum_{k=1}^{n} u_{x_k}x_k \). This is also written as \( \nabla^2 u \). The symbol, \( \Delta \) or \( \nabla^2 \) is called the Laplacian. When \( \Delta u = 0 \) the function, \( u \) is called harmonic. Laplace’s equation is \( \Delta u = 0 \). The heat equation is \( u_t - \Delta u = 0 \) and the wave equation is \( u_{tt} - \Delta u = 0 \).

**Example 16.5.2** Find the Laplacian of \( u(x, y) = x^2 - y^2 \).

\[ u_{xx} = 2 \text{ while } u_{yy} = -2. \text{ Therefore, } \Delta u = u_{xx} + u_{yy} = 2 - 2 = 0. \text{ Thus this function is harmonic.} \]

**Example 16.5.3** Find \( u_t - \Delta u \) where \( u(t, x, y) = e^{-t} \cos x \).

In this case, \( u_t = -e^{-t} \cos x \) while \( u_{yy} = 0 \) and \( u_{xx} = -e^{-t} \cos x \) therefore, \( u_t - \Delta u = 0 \) and so \( u \) solves the heat equation.

**Example 16.5.4** Let \( u(t, x) = \sin t \cos x \). Find \( u_{tt} - \Delta u \).

In this case, \( u_{tt} = -\sin t \cos x \) while \( \Delta u = -\sin t \cos x \). Therefore, \( u \) is a solution of the wave equation.

16.6 Exercises

1. Find the directional derivative of \( f(x, y, z) = x^2y + z^4 \) in the direction of the vector, \((1, 3, -1)\) when \((x, y, z) = (1, 1, 1)\).

2. Find the directional derivative of \( f(x, y, z) = \sin(x + y^2) + z \) in the direction of the vector, \((1, 2, -1)\) when \((x, y, z) = (1, 1, 1)\).

3. Find the directional derivative of \( f(x, y, z) = \ln(x + y^2) + z^2 \) in the direction of the vector, \((1, 1, -1)\) when \((x, y, z) = (1, 1, 1)\).

4. Find the largest value of the directional derivative of \( f(x, y, z) = \ln(x + y^2) + z^2 \) at the point \((1, 1, 1)\).

5. Find the smallest value of the directional derivative of \( f(x, y, z) = x \sin(4xy^2) + z^2 \) at the point \((1, 1, 1)\).

6. An ant falls to the top of a stove having temperature \( T(x, y) = x^2 \sin(x + y) \) at the point \((2, 3)\). In what direction should the ant go to minimize the temperature? In what direction should he go to maximize the temperature?

7. Find the partial derivative with respect to \( y \) of the function

\[
 f(x, y, z, w) = (y^2, z^2 \sin(xy), z^3x)^T.
\]

\(^3\)Laplace was a great physicist of the 1700’s. He made fundamental contributions to mechanics and astronomy.
8. Find the partial derivative with respect to $x$ of the function

$$f(x, y, z, w) = \begin{pmatrix} wx, zx \sin (xy), z^3x \end{pmatrix}^T.$$ 

9. Find $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, and $\frac{\partial f}{\partial z}$ for $f =$

(a) $x^2y + \cos (xy) + z^3y$
(b) $e^{x^2-y^2}z \sin (x + y)$
(c) $z^2 \sin^3 (e^{x^2+y})$
(d) $x^2 \cos (\sin (z^2 + y^2))$
(e) $xy^2z$

10. Suppose

$$f(x, y) = \begin{cases} 2xy + 12yz + 18x^2 + 36y^3 + \sin(x^2) + \tan(3y^3) & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}.$$ 

Find $\frac{\partial f}{\partial x}(0, 0)$ and $\frac{\partial f}{\partial y}(0, 0)$.

11. Why must the vector in the definition of the directional derivative be a unit vector? **Hint:** Suppose not. Would the directional derivative be a correct manifestation of steepness?

12. Find $f_x, f_y, f_z, f_{xy}, f_{xz}, f_{yz}, f_{xy}, f_{yz}$ for the following and form a conjecture about the mixed partial derivatives.

(a) $x^2y^3z^4 + \sin (xyz)$
(b) $\sin (xyz) + x^2y^2z$
(c) $z \ln |x^2 + y^2 + 1|$
(d) $e^{x^2+y^2+z^2}$
(e) $\tan (xyz)$

13. Suppose $f : U \to \mathbb{R}$ 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 $f_{x_i}(x) = 0$ for each $x_i$. **Hint:** This is just a repeat of the usual one variable theorem seen in beginning calculus. You just do this one variable argument for each variable to get the conclusion.

14. As an important application of Problem 13 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 $\{(t_i, x_i)\}_{i=1}^{n}$ 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 $b$, $\sum_{i=1}^{n} (at_i + b - x_i)^2$ is as small as possible. In other words, you want to minimize the function of two variables, $f(a, b) = \sum_{i=1}^{n} (at_i + b - x_i)^2$. 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.

15. Show that if $v(x, y) = u(\alpha x, \beta y)$, then $v_x = \alpha u_x$ and $v_y = \beta u_y$. State and prove a generalization to any number of variables.
16. Let \( f \) be a function which has continuous derivatives. Show \( u(t, x) = f(x - ct) \) solves the wave equation, \( u_{tt} - c^2 \Delta u = 0 \). What about \( u(x, t) = f(x + ct) \)?

17. Determine which of the following functions satisfy Laplace’s equation.

(a) \( x^3 - 3xy^2 \)
(b) \( 3x^2y - y^3 \)
(c) \( x^3 - 3xy^2 + 2x^2 - 2y^2 \)
(d) \( 3x^2y - y^3 + 4xy \)
(e) \( 3x^2 - y^3 + 4xy \)
(f) \( 3x^2y - y^3 + 4y \)
(g) \( x^3 - 3x^2y^2 + 2x^2 - 2y^2 \)

18. Show that if \( \Delta u = \lambda u \), then \( e^{\lambda t}u \) solves the heat equation, \( u_t - \Delta u = 0 \).

19. Show that if \( a, b \) are scalars and \( u, v \) are functions which satisfy Laplace’s equation then \( au + bv \) also satisfies Laplace’s equation. Verify a similar statement for the heat and wave equations.
The Derivative Of A Function Of Many Variables

17.0.1 Outcomes

1. Define differentiability and explain what the derivative is for a function of \( n \) variables.

2. Describe the relation between existence of partial derivatives, continuity, and differentiability.

3. Give examples of functions which have partial derivatives but are not continuous, examples of functions which are differentiable but not \( C^1 \), and examples of functions which are continuous without having partial derivatives.

4. Evaluate derivatives of composite functions using the chain rule.

5. Solve related rates problems using the chain rule.

17.1 The Derivative Of Functions Of One Variable

First we review the notion of the derivative of a function of one variable.

Observation 17.1.1 Suppose a function, \( f \) of one variable has a derivative at \( x \). Then

\[
\lim_{h \to 0} \frac{|f(x+h) - f(x) - f'(x)h|}{|h|} = 0.
\]

This observation follows from the definition of the derivative of a function of one variable, namely

\[
f'(x) \equiv \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.
\]

Definition 17.1.2 A vector valued function of a vector, \( \mathbf{v} \) is called \( o(\mathbf{v}) \) if

\[
\lim_{|\mathbf{v}| \to 0} \frac{o(\mathbf{v})}{|\mathbf{v}|} = 0. \tag{17.1}
\]

Thus the function \( f(x+h) - f(x) - f'(x)h \) is \( o(h) \). When we say a function is \( o(h) \), it is used like an adjective. It is like saying the function is white or black or green or fat or thin. The term is used very imprecisely. Thus

\[
o(\mathbf{v}) = o(\mathbf{v}) + o(\mathbf{v}), \quad o(\mathbf{v}) = 45o(\mathbf{v}), \quad o(\mathbf{v}) = o(\mathbf{v}) - o(\mathbf{v}), \quad etc.
\]
When you add two functions with the property of the above definition, you get another one having that same property. When you multiply by 45 the property is also retained as it is when you subtract two such functions. How could something so sloppy be useful? The notation is useful precisely because it prevents you from obsessing over things which are not relevant and should be ignored.

**Theorem 17.1.3** Let \( f : (a, b) \to \mathbb{R} \) be a function of one variable. Then \( f' (x) \) exists if and only if

\[
f(x + h) - f(x) = ph + o(h)
\]  

(17.2)

In this case, \( p = f'(x) \).

**Proof:** From the above observation it follows that if \( f'(x) \) does exist, then (17.2) holds. Suppose then that (17.2) is true. Then

\[
\frac{f(x + h) - f(x)}{h} = p + o(h)
\]

Taking a limit, you see that

\[
p = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}
\]

and that in fact this limit exists which shows that \( p = f'(x) \). This proves the theorem.

This theorem shows that one way to define \( f'(x) \) is as the number, \( p \), if there is one which has the property that

\[
f(x + h) = f(x) + ph + o(h).
\]

You should think of \( p \) as the linear transformation resulting from multiplication by the \( 1 \times 1 \) matrix, \( (p) \).

**Example 17.1.4** Let \( f(x) = x^3 \). Find \( f'(x) \).

\[
f(x + h) = (x + h)^3 = x^3 + 3x^2h + 3xh^2 + h^3 = f(x) + 3x^2h + (3xh + h^2)h.
\]

Since \( (3xh + h^2)h = o(h) \), it follows \( f'(x) = 3x^2 \).

**Example 17.1.5** Let \( f(x) = \sin(x) \). Find \( f'(x) \).

\[
f(x + h) - f(x) = \sin(x + h) - \sin(x) = \sin(x)\cos(h) + \cos(x)\sin(h) - \sin(x)
\]

\[
= \cos(x)\sin(h) + \sin(x)\frac{(\cos(h) - 1)}{h}h
\]

\[
= \cos(x)h + \cos(x)\frac{(\sin(h) - h)}{h}h + \sin(x)\frac{(\cos(h) - 1)}{h}h.
\]

Now

\[
\cos(x)\frac{(\sin(h) - h)}{h}h + \sin(x)\frac{(\cos(h) - 1)}{h}h = o(h).
\]

(17.3)

Remember the fundamental limits which allowed you to find the derivative of \( \sin(x) \) were

\[
\lim_{h \to 0} \frac{\sin(h)}{h} = 1, \quad \lim_{h \to 0} \frac{\cos(h) - 1}{h} = 0.
\]

(17.4)

These same limits are what is needed to verify (17.3).
17.2 The Derivative Of Functions Of Many Variables

This way of thinking about the derivative is exactly what is needed to define the derivative of a function of \( n \) variables.

**Definition 17.2.1** Let \( f : U \rightarrow \mathbb{R}^p \) where \( U \) is an open set in \( \mathbb{R}^n \) for \( n, p \geq 1 \) and let \( x \in U \) be given. Then \( f \) is defined to be differentiable at \( x \in U \) if and only if there exist column vectors, \( v_i \), such that for \( h = (h_1, \ldots, h_n)^T \),

\[
f(x + h) = f(x) + \sum_{i=1}^{n} v_i h_i + o(h).
\]  

(17.5)

The derivative of the function, \( f \), denoted by \( Df(x) \), is the linear transformation defined by multiplying by the matrix whose columns are the \( p \times 1 \) vectors, \( v_i \). Thus if \( w \) is a vector in \( \mathbb{R}^n \),

\[
Df(x) w = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix} w.
\]

It is common to think of this matrix as the derivative but strictly speaking, this is incorrect. The derivative is a linear transformation determined by multiplication by this matrix, called the standard matrix because it is based on the standard basis vectors for \( \mathbb{R}^n \). The subtle issues involved in a thorough exploration of this issue will be avoided for now. It will be fine to think of the above matrix as the derivative. Other notations which are often used for this matrix or the linear transformation are \( f'(x), J(x) \), and even \( \frac{\partial f}{\partial x} \) or \( \frac{df}{dx} \).

**Theorem 17.2.2** Suppose \( f \) is as given above in (17.5). Then

\[
v_k = \lim_{h \to 0} \frac{f(x+h e_k) - f(x)}{h} = \frac{\partial f}{\partial x_k}(x),
\]

the \( k \)th partial derivative.

**Proof:** Let \( h = (0, \ldots, h, 0, \ldots, 0)^T = h e_k \) where the \( h \) is in the \( k \)th slot. Then (17.5) reduces to

\[
f(x+h) = f(x) + v_k h + o(h).
\]

Therefore, dividing by \( h \)

\[
\frac{f(x+h e_k) - f(x)}{h} = v_k + \frac{o(h)}{h}
\]

and taking the limit,

\[
\lim_{h \to 0} \frac{f(x+h e_k) - f(x)}{h} = \lim_{h \to 0} \left( v_k + \frac{o(h)}{h} \right) = v_k
\]

and so, the above limit exists. This proves the theorem.

Let \( f : U \rightarrow \mathbb{R}^p \) where \( U \) is an open subset of \( \mathbb{R}^p \) and \( f \) is differentiable. It was just shown

\[
f(x + v) = f(x) + \sum_{j=1}^{p} \frac{\partial f(x)}{\partial x_j} v_j + o(v).
\]

Taking the \( i \)th coordinate of the above equation yields

\[
f_i(x + v) = f_i(x) + \sum_{j=1}^{p} \frac{\partial f_i(x)}{\partial x_j} v_j + o(v)
\]
and it follows that the term with a sum is nothing more than the \(i^{th}\) component of \(J(x)\) \(v\) where \(J(x)\) is the \(q \times p\) matrix,

\[
\begin{pmatrix}
\frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_p} \\
\frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_p} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial f_q}{\partial x_1} & \frac{\partial f_q}{\partial x_2} & \cdots & \frac{\partial f_q}{\partial x_p}
\end{pmatrix}.
\]

This gives the form of the matrix which defines the linear transformation, \(Df(x)\). Thus

\[
f(x + v) = f(x) + J(x)v + o(v) \tag{17.6}
\]

and to reiterate, the linear transformation which results by multiplication by this \(q \times p\) matrix is known as the derivative.

Sometimes we write \(x, y, z\) instead of \(x_1, x_2, \) and \(x_3\). This is to save on notation and is easier to write and to look at although it lacks generality. When this is done it is understood that \(x = x_1, y = x_2,\) and \(z = x_3\). Thus the derivative is the linear transformation determined by

\[
\begin{pmatrix}
f_{1x} & f_{1y} & f_{1z} \\
f_{2x} & f_{2y} & f_{2z} \\
f_{3x} & f_{3y} & f_{3z}
\end{pmatrix}.
\]

Example 17.2.3 Let \(A\) be a constant \(m \times n\) matrix and consider \(f(x) = Ax\). Find \(Df(x)\) if it exists.

\[
f(x + h) - f(x) = A(x + h) - A(x) = Ah + o(h).
\]

In fact in this case, \(o(h) = 0\). Therefore, \(Df(x) = A\). Note that this looks the same as the case in one variable, \(f(x) = ax\).

17.3 \(C^1\) Functions

Given a function of many variables, how can you tell if it is differentiable? Sometimes you have to go directly to the definition and verify it is differentiable from the definition. For example, you may have seen the following important example in one variable calculus.

Example 17.3.1 Let \(f(x) = \begin{cases} x^2 \sin \left(\frac{1}{x}\right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}\). Find \(Df(0)\).

\[
f(h) - f(0) = 0h + h^2 \sin \left(\frac{1}{h}\right) = o(h) \text{ and so } Df(0) = 0.
\]

If you find the derivative for \(x \neq 0\), it is totally useless information if what you want is \(Df(0)\). This is because the derivative, turns out to be discontinuous. Try it. Find the derivative for \(x \neq 0\) and try to obtain \(Df(0)\) from it. You see, in this example you had to revert to the definition to find the derivative.

It isn't really too hard to use the definition even for more ordinary examples.

Example 17.3.2 Let \(f(x, y) = \begin{pmatrix} x^2y + y^3 \\ y^3x \end{pmatrix}\). Find \(Df(1, 2)\).
First of all note that the thing you are after is a $2 \times 2$ matrix.

$$f(1, 2) = \begin{pmatrix} 6 \\ 8 \end{pmatrix}.$$  

Then

$$f(1 + h_1, 2 + h_2) - f(1, 2)$$

$$= \begin{pmatrix} (1 + h_1)^2 (2 + h_2) + (2 + h_2)^2 \\ (2 + h_2)^3 (1 + h_1) \end{pmatrix} - \begin{pmatrix} 6 \\ 8 \end{pmatrix}$$

$$= \begin{pmatrix} 5h_2 + 4h_1 + 2h_1h_2 + 2h_2^2 + h_1^2h_2 + h_2^3 \\ 8h_1 + 12h_2 + 12h_1h_2 + 6h_2^2 + 6h_2h_1 + h_2^2 + h_1^2 + h_2h_1 \end{pmatrix}$$

$$= \begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \begin{pmatrix} 2h_1h_2 + 2h_1^2 + h_1^2h_2 + h_2^3 \\ 12h_1h_2 + 6h_2^2 + 6h_2h_1 + h_2^2 + h_1^2 + h_2h_1 \end{pmatrix}$$

$$= \begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + o(h).$$

Therefore, the standard matrix of the derivative is $\begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix}$.

Most of the time, there is an easier way to conclude a derivative exists and to find it. It involves the notion of a $C^1$ function.

**Definition 17.3.3** When $f : U \to \mathbb{R}^p$ for $U$ an open subset of $\mathbb{R}^n$ and the vector valued functions $\frac{\partial f}{\partial x_i}$ are all continuous, (equivalently each $\frac{\partial f}{\partial x_i}$ is continuous), the function is said to be $C^1 (U)$. If all the partial derivatives up to order $k$ exist and are continuous, then the function is said to be $C^k$.

It turns out that for a $C^1$ function, all you have to do is write the matrix described in Theorem 17.2.2 and this will be the derivative. There is no question of existence for the derivative for such functions. This is the importance of the next few theorems.

**Theorem 17.3.4** Let $U$ be an open subset of $\mathbb{R}^2$ and suppose $f : U \to \mathbb{R}$ has the property that the partial derivatives $f_x$ and $f_y$ exist for $(x, y) \in U$ and are continuous at the point $(x_0, y_0)$. Then

$$f( (x_0, y_0) + (v_1, v_2) ) = f( x_0, y_0 ) + \frac{\partial f}{\partial x} (x_0, y_0) v_1 + \frac{\partial f}{\partial y} (x_0, y_0) v_2 + o(v).$$

That is, $f$ is differentiable.

**Proof:**

$$f( (x_0, y_0) + (v_1, v_2) ) - \left( f( x_0, y_0 ) + \frac{\partial f}{\partial x} (x_0, y_0) v_1 + \frac{\partial f}{\partial y} (x_0, y_0) v_2 \right)$$

$$= \left( f( x_0 + v_1, y_0 + v_2 ) - f( x_0, y_0 ) \right) - \left( \frac{\partial f}{\partial x} (x_0, y_0) v_1 + \frac{\partial f}{\partial y} (x_0, y_0) v_2 \right)$$

changes only in first component

changes only in second component

$$= \begin{pmatrix} f(x_0 + v_1, y_0 + v_2) - f(x_0, y_0 + v_2) + f(x_0, y_0 + v_2) - f(x_0, y_0) \end{pmatrix}$$
Let each component function, holds for each of the components of $x$.

Example 17.3.6

Define $f : U \rightarrow \mathbb{R}^q$.

By the mean value theorem, there exist numbers $s$ and $t$ in $[0, 1]$ such that this equals

$$
\frac{\partial f}{\partial x} (x_0 + sv_1, y_0 + sv_2) v_1 + \frac{\partial f}{\partial y} (x_0, y_0) v_2
$$

Therefore, letting $o(v)$ denote the expression in (17.7), and noticing that $|v_1|$ and $|v_2|$ are both no larger than $|v|$,.

It follows

$$
\frac{|o(v)|}{|v|} \leq \left( \left| \frac{\partial f}{\partial x} (x_0 + tv_1, y_0 + v_2) - \frac{\partial f}{\partial x} (x_0, y_0) \right| + \left| \frac{\partial f}{\partial y} (x_0, y_0 + sv_2) - \frac{\partial f}{\partial y} (x_0, y_0) \right| \right) |v|.
$$

Therefore, $\lim_{v \to 0} \frac{|o(v)|}{|v|} = 0$ because of the assumption that $f_x$ and $f_y$ are continuous at the point $(x_0, y_0)$ and this proves the theorem.

Having proved a theorem for scalar valued functions, one for vector valued functions follows immediately.

Theorem 17.3.5 Let $U$ be an open subset of $\mathbb{R}^p$ for $p \geq 1$ and suppose $f : U \rightarrow \mathbb{R}^q$ has the property that each component function, $f_i$, is differentiable at $x_0$. Then $f$ is differentiable at $x_0$.

Proof: Let $f(x) \equiv (f_1(x), \ldots, f_q(x))^T$. From the assumption each component function is differentiable, the following holds for each $k = 1, \ldots, q$.

$$
f_k (x_0 + v) = f_k (x_0) + \sum_{i=1}^p \frac{\partial f_k}{\partial x_i} (x_0) v_i + o_k (v).
$$

Define $o(v) \equiv (o_1(v), \ldots, o_q(v))^T$. Then (17.1) on Page 301 holds for $o(v)$ because it holds for each of the components of $o(v)$. The above equation is then equivalent to

$$
f (x_0 + v) = f (x_0) + \sum_{i=1}^p \frac{\partial f}{\partial x_i} (x_0) v_i + o (v)
$$

and so $f$ is differentiable at $x_0$.

Here is an example to illustrate.

Example 17.3.6 Let $f(x, y) = \left( \begin{array}{c} x^2 y + y^2 \\ y^3 x \end{array} \right)$. Find $Df(x, y)$. 

From Theorem 17.3.4 this function is differentiable because all possible partial derivatives are continuous. Thus

\[ Df(x, y) = \begin{pmatrix} 2xy & x^2 + 2y \\ y^3 & 3y^2x \end{pmatrix}. \]

In particular,

\[ Df(1, 2) = \begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix}. \]

Not surprisingly, the above theorem has an extension to more variables. First this is illustrated with an example.

**Example 17.3.7** Let \( f(x_1, x_2, x_3) = \begin{pmatrix} x_1^2 x_2 + x_2^2 \\ 2x_1 x_2 x_3 \cos (x_1 x_2 x_3) + x_1 x_3 \cos (x_1 x_2 x_3) \\ x_1 x_2 \cos (x_1 x_2 x_3) \end{pmatrix} \). Find \( Df(x_1, x_2, x_3) \).

All possible partial derivatives are continuous so the function is differentiable. The matrix for this derivative is therefore the following \( 3 \times 3 \) matrix

\[ \begin{pmatrix} 2x_1 x_2 & x_1^2 + 2x_2 & 0 \\ x_2 & 1 & 0 \\ x_2 x_3 \cos (x_1 x_2 x_3) & x_1 x_3 \cos (x_1 x_2 x_3) & x_1 x_2 \cos (x_1 x_2 x_3) \end{pmatrix}. \]

The following theorem is the general result.

**Theorem 17.3.8** Let \( U \) be an open subset of \( \mathbb{R}^p \) for \( p \geq 1 \) and suppose \( f: U \rightarrow \mathbb{R} \) has the property that the partial derivatives \( f_{x_i} \) exist for all \( x \in U \) and are continuous at the point \( x_0 \in U \). Then

\[ f(x_0 + v) = f(x_0) + \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x_0) v_i + o(v). \]

That is, \( f \) is differentiable at \( x_0 \) and the derivative of \( f \) equals the linear transformation obtained by multiplying by the \( 1 \times p \) matrix,

\[ \left( \frac{\partial f}{\partial x_1}(x_0), \ldots, \frac{\partial f}{\partial x_p}(x_0) \right). \]

**Proof:** The proof is similar to the case of two variables. Letting \( v = (v_1, \ldots, v_p)^T \), denote by \( \theta^i v \) the vector \((0, \ldots, 0, v_i, v_{i+1}, \ldots, v_p)^T\). Thus \( \theta_0 v = v, \theta_{p-1}(v) = (0, \ldots, 0, v_p)^T \), and \( \theta_p v = 0 \). Now

\[ f(x_0 + v) - f(x_0) = \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x_0) v_i \]

(17.8)

= \sum_{i=1}^{p} \left( f(x_0 + \theta_{i-1} v) - f(x_0 + \theta_i v) \right) - \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x_0) v_i

Now by the mean value theorem there exist numbers \( s_i \in (0, 1) \) such that the above expression equals

\[ = \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x_0 + \theta_i v + s_i v_i) - \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x_0) v_i. \]
and so letting $o(v)$ equal the expression in (17.8),

$$|o(v)| \leq \sum_{i=1}^{p} \left| \frac{\partial f}{\partial x_i} (x_0 + \theta_i v + s_i v_i) - \frac{\partial f}{\partial x_i} (x_0) \right| |v_i|$$

$$\leq \sum_{i=1}^{p} \left| \frac{\partial f}{\partial x_i} (x_0 + \theta_i v + s_i v_i) - \frac{\partial f}{\partial x_i} (x_0) \right| |v|$$

and so

$$\lim_{v \to 0} \frac{|o(v)|}{|v|} \leq \lim_{v \to 0} \sum_{i=1}^{p} \left| \frac{\partial f}{\partial x_i} (x_0 + \theta_i v + s_i v_i) - \frac{\partial f}{\partial x_i} (x_0) \right| = 0$$

because of continuity of the $f_x$, at $x_0$. This proves the theorem.

Letting $x - x_0 = v$,

$$f(x) = f(x_0) + \sum_{i=1}^{p} \frac{\partial f}{\partial x_i} (x_0) (x_i - x_{0i}) + o(v)$$

$$= f(x_0) + \sum_{i=1}^{p} \frac{\partial f}{\partial x_i} (x_0) v_i + o(v).$$

**Example 17.3.9** Suppose $f(x, y, z) = xy + z^2$. Find $Df(1, 2, 3)$.

Taking the partial derivatives of $f$, $f_x = y$, $f_y = x$, $f_z = 2z$. These are all continuous. Therefore, the function has a derivative and $f_x(1, 2, 3) = 1$, $f_y(1, 2, 3) = 2$, and $f_z(1, 2, 3) = 6$. Therefore, $Df(1, 2, 3)$ is given by

$$Df(1, 2, 3) = (1, 2, 6).$$

Also, for $(x, y, z)$ close to $(1, 2, 3)$,

$$f(x, y, z) \approx f(1, 2, 3) + 1(x - 1) + 2(y - 2) + 6(z - 3)$$

$$= 11 + 1(x - 1) + 2(y - 2) + 6(z - 3) = -12 + x + 2y + 6z$$

In the case where $f$ has values in $\mathbb{R}^q$ rather than $\mathbb{R}$, is there a similar theorem about differentiability of a $C^1$ function?

**Theorem 17.3.10** Let $U$ be an open subset of $\mathbb{R}^p$ for $p \geq 1$ and suppose $f : U \to \mathbb{R}^q$ has the property that the partial derivatives $f_{x_i}$ exist for all $x \in U$ and are continuous at the point $x_0 \in U$, then

$$f(x_0 + v) = f(x_0) + \sum_{i=1}^{p} \frac{\partial f}{\partial x_i} (x_0) v_i + o(v) \quad (17.9)$$

and so $f$ is differentiable at $x_0$.

**Proof:** This follows from Theorem 17.3.5.

When a function is differentiable at $x_0$ it follows the function must be continuous there. This is the content of the following important lemma.

**Lemma 17.3.11** Let $f : U \to \mathbb{R}^q$ where $U$ is an open subset of $\mathbb{R}^p$. If $f$ is differentiable, then $f$ is continuous at $x_0$. Furthermore, if $C = \max \left\{ \left| \frac{\partial f}{\partial x_i} (x_0) \right|, i = 1, \cdots, p \right\}$, then whenever $|x - x_0|$ is small enough,

$$|f(x) - f(x_0)| \leq (Cp + 1)|x - x_0| \quad (17.10)$$
**Proof:** Suppose $f$ is differentiable. Since $o(v)$ satisfies (17.1), there exists $\delta_1 > 0$ such that if $|x - x_0| < \delta_1$, then $|o(x - x_0)| < |x - x_0|$. But also, by the triangle inequality, Corollary 6.1.5 on Page 116,

$$\left| \sum_{i=1}^{p} \frac{\partial f}{\partial x_i} (x_0) (x_i - x_{0i}) \right| \leq C \sum_{i=1}^{p} |x_i - x_{0i}| \leq Cp |x - x_0|$$

Therefore, if $|x - x_0| < \delta_1$,

$$|f(x) - f(x_0)| \leq \left| \sum_{i=1}^{p} \frac{\partial f}{\partial x_i} (x_0) (x_i - x_{0i}) \right| + |x - x_0| \leq (Cp + 1) |x - x_0|$$

which verifies (17.10). Now letting $\varepsilon > 0$ be given, let $\delta = \min \left( \delta_1, \frac{\varepsilon}{Cp+1} \right)$. Then for $|x - x_0| < \delta$,

$$|f(x) - f(x_0)| < (Cp + 1) |x - x_0| < (Cp + 1) \frac{\varepsilon}{Cp + 1} = \varepsilon$$

showing $f$ is continuous at $x_0$.

There have been quite a few terms defined. First there was the concept of continuity. Next the concept of partial or directional derivative. Next there was the concept of differentiability and the derivative being a linear transformation determined by a certain matrix. Finally, it was shown that if a function is $C^1$, then it has a derivative. To give a rough idea of the relationships of these topics, here is a picture.

---

You might ask whether there are examples of functions which are differentiable but not $C^1$. Of course there are. In fact, Example 17.3.1 is just such an example as explained earlier. Then you should verify that $f'(x)$ exists for all $x \in \mathbb{R}$ but $f'$ fails to be continuous at $x = 0$. Thus the function is differentiable at every point of $\mathbb{R}$ but fails to be $C^1$ at every point of $\mathbb{R}$.

### 17.3.1 Approximation With A Tangent Plane

In the case where $f$ is a scalar valued function of two variables, the geometric significance of the derivative can be exhibited in the following picture. Writing $v \equiv (x - x_0, y - y_0)$, the notion of differentiability at $(x_0, y_0)$ reduces to

$$f(x, y) = f(x_0, y_0) + \frac{\partial f}{\partial x} (x_0, y_0) (x - x_0) + \frac{\partial f}{\partial y} (x_0, y_0) (y - y_0) + o(v)$$
The right side of the above, \( f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0) = z \) is the equation of a plane approximating the graph of \( z = f(x, y) \) for \((x, y)\) near \((x_0, y_0)\). Saying that the function is differentiable at \((x_0, y_0)\) amounts to saying that the approximation delivered by this plane is very good if both \(|x - x_0|\) and \(|y - y_0|\) are small.

**Example 17.3.12** Suppose \( f(x, y) = \sqrt{xy} \). Find the approximate change in \( f \) if \( x \) goes from 1 to 1.01 and \( y \) goes from 4 to 3.99.

We can do this by noting that

\[
\begin{align*}
 f(1.01, 3.99) - f(1, 4) & \approx f_x(1, 2)(.01) + f_y(1, 2)(-.01) \\
 & = 1(.01) + \frac{1}{4}(-.01) = 7.5 \times 10^{-3}.
\end{align*}
\]

Of course the exact value would be

\[
\sqrt{(1.01)(3.99)} - \sqrt{4} = 7.4610831 \times 10^{-3}.
\]

### 17.4 The Chain Rule

As in the case of a function of one variable, it is important to consider the derivative of a composition of two functions. The proof of the chain rule depends on the following fundamental lemma.

**Lemma 17.4.1** Let \( g : U \to \mathbb{R}^p \) where \( U \) is an open set in \( \mathbb{R}^n \) and suppose \( g \) has a derivative at \( x \in U \). Then 

\[
o(g(x + v) - g(x)) = o(v).
\]

**Proof:** It is necessary to show

\[
\lim_{v \to 0} \frac{|o(g(x + v) - g(x))|}{|v|} = 0. \quad (17.11)
\]

From Lemma 17.3.11, there exists \( \delta > 0 \) such that if \(|v| < \delta\), then

\[
|g(x + v) - g(x)| \leq (Cn + 1)|v|. \quad (17.12)
\]

Now let \( \varepsilon > 0 \) be given. There exists \( \eta > 0 \) such that if \(|g(x + v) - g(x)| < \eta\), then

\[
|o(g(x + v) - g(x))| < \left( \frac{\varepsilon}{Cn + 1} \right) |g(x + v) - g(x)| \quad (17.13)
\]

Let \(|v| < \min\left( \delta, \frac{\eta}{Cn + 1} \right)\). For such \( v \), \(|g(x + v) - g(x)| \leq \eta\), which implies

\[
|o(g(x + v) - g(x))| < \left( \frac{\varepsilon}{Cn + 1} \right) |g(x + v) - g(x)| < \left( \frac{\varepsilon}{Cn + 1} \right)(Cn + 1)|v|.
\]
and so

$$\frac{|o(g(x + v) - g(x))|}{|v|} < \varepsilon$$

which establishes (17.11). This proves the lemma.

Recall the notation $f \circ g(x) \equiv f(g(x))$. Thus $f \circ g$ is the name of a function and this function is defined by what was just written. The following theorem is known as the chain rule.

**Theorem 17.4.2 (Chain rule)** Let $U$ be an open set in $\mathbb{R}^n$, let $V$ be an open set in $\mathbb{R}^p$, let $g : U \to \mathbb{R}^p$ be such that $g(U) \subseteq V$, and let $f : V \to \mathbb{R}^q$. Suppose $Dg(x)$ exists for some $x \in U$ and that $Df(g(x))$ exists. Then $D(f \circ g)(x)$ exists and furthermore,

$$D(f \circ g)(x) = Df(g(x)) \cdot Dg(x).$$

(17.14) In particular,

$$\frac{\partial (f \circ g)(x)}{\partial x_j} = \sum_{i=1}^p \frac{\partial f(g(x))}{\partial y_i} \frac{\partial y_i}{\partial x_j}.$$  

(17.15)

**Proof:** From the assumption that $Df(g(x))$ exists,

$$f(g(x + v)) = f(g(x)) + \sum_{i=1}^p \frac{\partial f(g(x))}{\partial y_i} (g_i(x + v) - g_i(x)) + o(g(x + v) - g(x))$$

which by Lemma 17.4.1 equals

$$(f \circ g)(x + v) = f(g(x + v)) = f(g(x)) + \sum_{i=1}^p \frac{\partial f(g(x))}{\partial y_i} (g_i(x + v) - g_i(x)) + o(v).$$

Now since $Dg(x)$ exists, the above becomes

$$(f \circ g)(x + v) = f(g(x)) + \sum_{i=1}^p \frac{\partial f(g(x))}{\partial y_i} \left( \sum_{j=1}^n \frac{\partial y_i}{\partial x_j} v_j + o(v) \right) + o(v)$$

$$= f(g(x)) + \sum_{i=1}^p \frac{\partial f(g(x))}{\partial y_i} \left( \sum_{j=1}^n \frac{\partial y_i}{\partial x_j} v_j \right) + \sum_{i=1}^p \frac{\partial f(g(x))}{\partial y_i} o(v) + o(v)$$

$$= (f \circ g)(x) + \sum_{j=1}^n \left( \sum_{i=1}^p \frac{\partial f(g(x))}{\partial y_i} \frac{\partial y_i}{\partial x_j} \right) v_j + o(v)$$

because $\sum_{i=1}^p \frac{\partial f(g(x))}{\partial y_j} o(v) + o(v) = o(v)$. This establishes (17.15) because of Theorem 17.2.2 on Page 303. Thus

$$D(f \circ g)(x)_{kj} = \sum_{i=1}^p \frac{\partial f_k(g(x))}{\partial y_i} \frac{\partial y_i}{\partial x_j}$$

and

$$= \sum_{i=1}^p \frac{\partial f(g(x))_{ki}}{\partial x_j} (Dg(x))_{ij}.$$
There is an easy way to remember this in terms of the repeated index summation convention presented earlier. Let \( y = g(x) \) and \( z = f(y) \). Then the above says

\[
\frac{\partial z}{\partial y_i} \frac{\partial y_i}{\partial x_k} = \frac{\partial z}{\partial x_k}
\]

Remember there is a sum on the repeated index.

**Example 17.4.3** Let \( f(u, v) = \sin(uv) \) and let \( u(x, y, t) = t \sin x + \cos y \) and \( v(x, y, t, s) = s \tan x + y^2 + ts \). Letting \( z = f(u, v) \) where \( u, v \) are as just described, find \( \frac{\partial z}{\partial t} \) and \( \frac{\partial z}{\partial x} \).

From the above,

\[
\frac{\partial z}{\partial t} = \frac{\partial z}{\partial u} \frac{\partial u}{\partial t} + \frac{\partial z}{\partial v} \frac{\partial v}{\partial t} = v \cos(\sin x) \sin x + u \cos(uv).
\]

Also,

\[
\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u} \frac{\partial u}{\partial x} + \frac{\partial z}{\partial v} \frac{\partial v}{\partial x} = v \cos(\sin x) \cos(x) + u \sec^2(x) \cos(uv).
\]

Clearly you can continue in this way taking partial derivatives with respect to any of the other variables.

**Example 17.4.4** Let \( f(u_1, u_2) = \left( \begin{array}{c} u_1^2 + u_2 \\ \sin(u_2) + u_1 \end{array} \right) \) and \( g(x_1, x_2, x_3) = \left( \begin{array}{c} u_1(x_1, x_2, x_3) \\ u_2(x_1, x_2, x_3) \end{array} \right) \) = \( \left( \begin{array}{c} x_1x_2 + x_3 \\ x_2^3 + x_1 \end{array} \right) \). Find \( D(f \circ g)(x_1, x_2, x_3) \).

To do this,

\[
Df(u_1, u_2) = \left( \begin{array}{c} 2u_1 \\ \cos u_2 \end{array} \right), \quad Dg(x_1, x_2, x_3) = \left( \begin{array}{ccc} x_2 & x_1 & 1 \\ x_2 & 2x_2 & 0 \end{array} \right)
\]

Then

\[
Df(g(x_1, x_2, x_3)) = \left( \begin{array}{c} 2(x_1x_2 + x_3) \\ \cos(x_2^3 + x_1) \end{array} \right)
\]

and so by the chain rule,

\[
D(f \circ g)(x_1, x_2, x_3) = \begin{pmatrix}
Df(g(x)) & Dg(x)
\end{pmatrix} = \left( \begin{array}{c c}
2(x_1x_2 + x_3) & x_2 \\
\cos(x_2^3 + x_1) & x_1 + 2x_2 \cos(x_2^3 + x_1) \\
\end{array} \right)
\left( \begin{array}{c}
x_2 \\
x_1 + 2x_2 \cos(x_2^3 + x_1)
\end{array} \right)
\]

Therefore, in particular,

\[
\frac{\partial f_1 \circ g}{\partial x_1}(x_1, x_2, x_3) = (2x_1x_2 + 2x_3)x_2 + 1,
\]

\[
\frac{\partial f_2 \circ g}{\partial x_3}(x_1, x_2, x_3) = 1, \quad \frac{\partial f_2 \circ g}{\partial x_2}(x_1, x_2, x_3) = x_1 + 2x_2 \left( \cos(x_2^3 + x_1) \right).
\]

etc.

**Example 17.4.5** Let \( f:U \to V \) where \( U \) and \( V \) are open sets in \( \mathbb{R}^n \) and \( f \) is one to one and onto. Suppose also that \( f \) and \( f^{-1} \) are both differentiable. How are \( DF^{-1} \) and \( DF \) related?
This can be done as follows. From the assumptions, \( x = f^{-1}(f(x)) \). Let \( Ix = x \). Then by Example 17.2.3 on Page 304 \( DI = I \). By the chain rule,

\[
I = DI = Df^{-1}(f(x))(Df(x)).
\]

Therefore,

\[
Df(x)^{-1} = Df^{-1}(f(x)).
\]

This is equivalent to

\[
Df(f^{-1}(y))^{-1} = 1/\left( f'\left( f^{-1}(y) \right) \right).
\]

or

\[
Df(x)^{-1} = Df^{-1}(y), \quad y = f(x),
\]

This is just like a similar situation for functions of one variable. Remember \((f^{-1})'\left( f(x) \right) = 1/f'(x)\). In terms of the repeated index summation convention, suppose \( y = f(x) \) so that \( x = f^{-1}(y) \). Then the above can be written as

\[
\delta_{ij} = \frac{\partial x_i}{\partial y_k} \frac{\partial y_k}{\partial x_j}(x).
\]

**Example 17.4.6** Recall spherical coordinates are given by

\[
x = \rho \sin \phi \cos \theta, \quad y = \rho \sin \phi \sin \theta, \quad z = \rho \cos \phi.
\]

If an object moves in three dimensions, describe its acceleration in terms of spherical coordinates and the vectors,

\[
e_\rho = (\sin \phi \cos \theta, \sin \phi \sin \theta, \cos \phi)^T,
\]

\[
e_\theta = (-\rho \sin \phi \sin \theta, \rho \sin \phi \cos \theta, 0)^T,
\]

and

\[
e_\phi = (\rho \cos \phi \cos \theta, \rho \cos \phi \sin \theta, -\rho \sin \phi)^T.
\]

Why these vectors? Note how they were obtained. Let

\[
r(\rho, \theta, \phi) = (\rho \sin \phi \cos \theta, \rho \sin \phi \sin \theta, \rho \cos \phi)^T
\]

and fix \( \phi \) and \( \theta \), letting only \( \rho \) change, this gives a curve in the direction of increasing \( \rho \). Thus it is a vector which points away from the origin. Letting only \( \phi \) change and fixing \( \theta \) and \( \rho \), this gives a vector which is tangent to the sphere of radius \( \rho \) and points South. Similarly, letting \( \theta \) change and fixing the other two gives a vector which points East and is tangent to the sphere of radius \( \rho \). It is thought by most people that we live on a large sphere. The model of a flat earth is not believed by anyone except perhaps for beginning physics students. Given we live on a sphere, what directions would be most meaningful? Wouldn’t it be the directions of the vectors just described?

Let \( r(t) \) denote the position vector of the object from the origin. Thus

\[
r(t) = \rho(t)e_\rho(t) = \left( (x(t), y(t), z(t))^T \right)
\]

Now this implies the velocity is

\[
r'(t) = \rho'(t)e_\rho(t) + \rho(t)e_\rho'(t).
\]

(17.16)
You see, \( \mathbf{e}_\rho = \mathbf{e}_\rho(\rho, \theta, \phi) \) where each of these variables is a function of \( t \).

\[
\frac{\partial \mathbf{e}_\rho}{\partial \phi} = (\cos \phi \cos \theta, \cos \phi \sin \theta, -\sin \phi)^T = \frac{1}{\rho} \mathbf{e}_\phi,
\]

\[
\frac{\partial \mathbf{e}_\rho}{\partial \theta} = (-\sin \phi \sin \theta, \sin \phi \cos \theta, 0)^T = \frac{1}{\rho} \mathbf{e}_\theta,
\]

and

\[
\frac{\partial \mathbf{e}_\rho}{\partial \rho} = 0.
\]

Therefore, by the chain rule,

\[
\frac{d \mathbf{e}_\rho}{dt} = \frac{\partial \mathbf{e}_\rho}{\partial \phi} \frac{d \phi}{dt} + \frac{\partial \mathbf{e}_\rho}{\partial \theta} \frac{d \theta}{dt} = \frac{1}{\rho} \mathbf{e}_\phi \frac{d \phi}{dt} + \frac{1}{\rho} \mathbf{e}_\theta \frac{d \theta}{dt}.
\]

By (17.16),

\[
r' = \rho' \mathbf{e}_\rho + \frac{d \phi}{dt} \mathbf{e}_\phi + \frac{d \theta}{dt} \mathbf{e}_\theta.
\]  

(17.17)

Now things get interesting. This must be differentiated with respect to \( t \). To do so,

\[
\frac{\partial \mathbf{e}_\theta}{\partial \theta} = \begin{pmatrix} -\rho \sin \phi \sin \theta & \rho \cos \phi \cos \theta & \sin \phi \cos \theta \\ \rho \sin \phi \cos \theta & \rho \cos \phi \sin \theta & \sin \phi \sin \theta \\ 0 & -\rho \sin \phi & \cos \phi \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} -\rho \sin \phi \cos \theta \\ -\rho \sin \phi \sin \theta \\ 0 \end{pmatrix}
\]

Using Cramer's rule, the solution is \( a = 0, b = -\cos \phi \sin \phi \), and \( c = -\rho \sin^2 \phi \). Thus

\[
\frac{\partial \mathbf{e}_\theta}{\partial \theta} = \begin{pmatrix} -\rho \sin \phi \cos \theta, -\rho \sin \phi \sin \theta, 0 \end{pmatrix}^T
\]

Also,

\[
\frac{\partial \mathbf{e}_\phi}{\partial \phi} = \begin{pmatrix} -\rho \cos \phi \sin \theta, \rho \cos \phi \cos \theta, 0 \end{pmatrix}^T = (\cot \phi) \mathbf{e}_\theta
\]

and

\[
\frac{\partial \mathbf{e}_\theta}{\partial \rho} = \begin{pmatrix} -\sin \phi \sin \theta, \sin \phi \cos \theta, 0 \end{pmatrix}^T = \frac{1}{\rho} \mathbf{e}_\theta.
\]

Now in (17.17) it is also necessary to consider \( \mathbf{e}_\phi \). 

\[
\frac{\partial \mathbf{e}_\phi}{\partial \phi} = \begin{pmatrix} -\rho \sin \phi \cos \theta, -\rho \sin \phi \sin \theta, -\rho \cos \phi \end{pmatrix}^T = -\rho \mathbf{e}_\rho
\]

\[
\frac{\partial \mathbf{e}_\phi}{\partial \theta} = \begin{pmatrix} -\rho \cos \phi \sin \theta, \rho \cos \phi \cos \theta, 0 \end{pmatrix}^T = (\cot \phi) \mathbf{e}_\theta
\]
and finally,
\[ \frac{\partial \mathbf{e}_\phi}{\partial \rho} = (\cos \phi \cos \theta, \cos \phi \sin \theta, -\sin \phi)^T = \frac{1}{\rho} \mathbf{e}_\rho. \]

With these formulas for various partial derivatives, the chain rule is used to obtain \( r'' \) which will yield a formula for the acceleration in terms of the spherical coordinates and these special vectors. By the chain rule,
\[
\frac{d}{dt}(\mathbf{e}_\rho) = \frac{\partial \mathbf{e}_\rho}{\partial \theta} \theta' + \frac{\partial \mathbf{e}_\rho}{\partial \phi} \phi' + \frac{\partial \mathbf{e}_\rho}{\partial \rho} \rho'
\]
\[
= \frac{\theta'}{\rho} \mathbf{e}_\theta + \frac{\phi'}{\rho} \mathbf{e}_\phi
\]

\[
\frac{d}{dt}(\mathbf{e}_\theta) = \frac{\partial \mathbf{e}_\theta}{\partial \theta} \theta' + \frac{\partial \mathbf{e}_\theta}{\partial \phi} \phi' + \frac{\partial \mathbf{e}_\theta}{\partial \rho} \rho'
\]
\[
= \theta' \left( (\cos \phi \sin \phi) \mathbf{e}_\phi + (\rho \sin^2 \phi) \mathbf{e}_\rho \right) + \phi' \left( \cot \phi \right) \mathbf{e}_\theta + \left( \frac{\rho'}{\rho} \mathbf{e}_\rho \right)
\]

\[
\frac{d}{dt}(\mathbf{e}_\phi) = \frac{\partial \mathbf{e}_\phi}{\partial \theta} \theta' + \frac{\partial \mathbf{e}_\phi}{\partial \phi} \phi' + \frac{\partial \mathbf{e}_\phi}{\partial \rho} \rho'
\]
\[
= \left( \theta' \cot \phi \right) \mathbf{e}_\theta + \phi' \left( -\rho \mathbf{e}_\rho \right) + \left( \frac{\rho'}{\rho} \mathbf{e}_\rho \right)
\]

By (17.17),
\[
r'' = \rho'' \mathbf{e}_\rho + \phi'' \mathbf{e}_\phi + \theta'' \mathbf{e}_\theta + \rho' \left( \mathbf{e}_\rho \right)' + \phi' \left( \mathbf{e}_\phi \right)' + \theta' \left( \mathbf{e}_\theta \right)'
\]

and from the above, this equals
\[
\rho'' \mathbf{e}_\rho + \phi'' \mathbf{e}_\phi + \theta'' \mathbf{e}_\theta + \rho' \left( \frac{\theta'}{\rho} \mathbf{e}_\theta + \frac{\phi'}{\rho} \mathbf{e}_\phi \right) + \\
\phi' \left( \left( \theta' \cot \phi \right) \mathbf{e}_\theta + \phi' \left( -\rho \mathbf{e}_\rho \right) + \left( \frac{\rho'}{\rho} \mathbf{e}_\rho \right) \right) + \\
\theta' \left( (\cos \phi \sin \phi) \mathbf{e}_\phi + (\rho \sin^2 \phi) \mathbf{e}_\rho \right) + \phi' \left( \cot \phi \right) \mathbf{e}_\theta + \left( \frac{\rho'}{\rho} \mathbf{e}_\rho \right)
\]

and now all that remains is to collect the terms. Thus \( r'' \) equals
\[
r'' = \left( \rho'' - \rho \left( \phi' \right)^2 - \rho \left( \theta' \right)^2 \sin^2 \phi \right) \mathbf{e}_\rho + \left( \phi'' + \frac{2\rho' \phi'}{\rho} - (\theta')^2 \cos \phi \sin \phi \right) \mathbf{e}_\phi + \\
+ \left( \theta'' + \frac{2\theta' \rho'}{\rho} + 2\phi' \theta' \cot \phi \right) \mathbf{e}_\theta
\]

and this gives the acceleration in spherical coordinates. Note the prominent role played by the chain rule. All of the above is done in books on mechanics for general curvilinear coordinate systems and in the more general context, special theorems are developed which make things go much faster but these theorems are all exercises in the chain rule.

As an example of how this could be used, consider a rocket. Suppose for simplicity that it experiences a force only in the direction of \( \mathbf{e}_\rho \), directly away from the earth. Of course this force produces a corresponding acceleration which can be computed as a function of time. As the fuel is burned, the rocket becomes less massive and so the acceleration will be an increasing function of \( t \). However, this would be a known function, say \( a(t) \). Suppose
you wanted to know the latitude and longitude of the rocket as a function of time. (There is no reason to think these will stay the same.) Then all that would be required would be to solve the system of differential equations,  

\[ \rho'' - \rho \left( \phi' \right)^2 - \rho \left( \theta' \right)^2 \sin^2 (\phi) = a(t), \]

\[ \phi'' + \frac{2 \rho' \phi'}{\rho} - \left( \theta' \right)^2 \cos \phi \sin \phi = 0, \]

\[ \theta'' + \frac{2 \theta' \rho'}{\rho} + 2 \phi' \theta' \cot (\phi) = 0, \]

along with initial conditions, \( \rho(0) = \rho_0 \) (the distance from the launch site to the center of the earth.), \( \phi'(0) = \phi_1 \) (the initial vertical component of velocity of the rocket, probably 0.), and then initial conditions for \( \phi, \phi', \theta, \theta' \). The initial value problems could then be solved numerically and you would know the distance from the center of the earth as a function of \( t \) along with \( \theta \) and \( \phi \). Thus you could predict where the booster shells would fall to earth so you would know where to look for them. Of course there are many variations of this. You might want to specify forces in the \( e_\theta \) and \( e_\phi \) direction as well and attempt to control the position of the rocket or rather its payload. The point is that if you are interested in doing all this in terms of \( \phi, \theta, \) and \( \rho \), the above shows how to do it systematically and you see it is all an exercise in using the chain rule. More could be said here involving moving coordinate systems and the Coriolis force. You really might want to do everything with respect to a coordinate system which is fixed with respect to the moving earth.

### 17.5 Lagrangian Mechanics*

A difficult and important problem is to come up with differential equations which model mechanical systems. Lagrange gave a way to do this. It will be presented here as a very interesting and important application of the chain rule. Lagrange developed this technique back in the 1700’s. The presentation here follows [11]. Assume \( N \) point masses, located at the points \( x_1, \cdots, x_N \) in \( \mathbb{R}^3 \) and let the mass of the \( \alpha \)th mass be \( m_\alpha \). Then according to Newton’s second law,

\[ m_\alpha \ddot{x}_\alpha = F_\alpha (x_\alpha, t). \tag{17.18} \]

The dependence of \( F_\alpha \) on the two indicated quantities is indicative of the situation where the force may change in time and position. Now define

\[ x \equiv (x_1, \cdots, x_N) \in \mathbb{R}^{3N} \]

and assume \( x \in M \) which is defined locally in the form \( x = G(q, t) \). Here \( q \in \mathbb{R}^m \) where typically \( m < 3N \) and \( G(\cdot, t) \) is a smooth one to one mapping from \( V \), an open subset of \( \mathbb{R}^m \) onto a set of points near \( x \) which are on \( M \). Also assume \( t \) is in an open subset of \( \mathbb{R} \). In what follows a dot over a variable will indicate a derivative taken with respect to time. Two dots will indicate the second derivative with respect to time, etc. Then define \( G_\alpha \) by

\[ x_\alpha = G_\alpha (q, t). \]

Using the summation convention and the chain rule,

\[ \frac{dx_\alpha}{dt} = \frac{\partial G_\alpha}{\partial q} \frac{dq}{dt} + \frac{\partial G_\alpha}{\partial t}. \]

---

1 You won’t be able to find the solution to equations like these in terms of simple functions. However, they can be solved numerically. This means you determine the value of the various variables for various values of \( t \) without finding a neat formula involving known functions for the solution. This sort of computation is usually done by a computer.
Therefore, the kinetic energy is of the form

\[
T = \sum_{\alpha=1}^{N} \frac{1}{2} m_{\alpha} \left( \frac{dx_{\alpha}}{dt} \cdot \frac{dx_{\alpha}}{dt} \right)
\]

\[
= \sum_{\alpha=1}^{N} \frac{1}{2} m_{\alpha} \left( \sum_{j} \frac{\partial G_{\alpha}}{\partial q^{j}} \frac{dq^{j}}{dt} + \frac{\partial G_{\alpha}}{\partial t} \right)
\]

\[
= \sum_{j,r} \frac{1}{2} \left[ \sum_{\alpha} m_{\alpha} \left( \frac{\partial G_{\alpha}}{\partial q^{j}} \cdot \frac{\partial G_{\alpha}}{\partial q^{r}} \right) \right] \dot{q}^{j} \dot{q}^{r} + \sum_{\alpha} \sum_{j} m_{\alpha} \left( \frac{\partial G_{\alpha}}{\partial q^{j}} \cdot \frac{\partial G_{\alpha}}{\partial t} \right) \dot{q}^{j}
\]

\[
+ \sum_{\alpha} \frac{1}{2} m_{\alpha} \left( \frac{\partial G_{\alpha}}{\partial t} \cdot \frac{\partial G_{\alpha}}{\partial t} \right)
\]

(17.19)

where in the last equation \( \dot{q}^{k} \) indicates \( \frac{dq^{k}}{dt} \). Therefore,

\[
\frac{\partial T}{\partial q^{k}} = \sum_{j=1}^{m} \sum_{\alpha=1}^{N} m_{\alpha} \left( \frac{\partial G_{\alpha}}{\partial q^{j}} \cdot \frac{\partial G_{\alpha}}{\partial q^{k}} \right) \dot{q}^{j} + \sum_{\alpha} m_{\alpha} \left( \frac{\partial G_{\alpha}}{\partial q^{k}} \cdot \frac{\partial G_{\alpha}}{\partial t} \right)
\]

\[
= \sum_{\alpha=1}^{N} \left( m_{\alpha} \frac{\partial G_{\alpha}}{\partial q^{k}} \cdot \sum_{j=1}^{m} \frac{\partial G_{\alpha}}{\partial q^{j}} \dot{q}^{j} \right) + \sum_{\alpha} m_{\alpha} \left( \frac{\partial G_{\alpha}}{\partial q^{k}} \cdot \frac{\partial G_{\alpha}}{\partial t} \right)
\]

\[
= \sum_{\alpha=1}^{N} \frac{\partial G_{\alpha}}{\partial q^{k}} \cdot m_{\alpha} x^{n}_{\alpha}
\]

Now using the chain rule and product rule again, along with Newton's second law,

\[
\frac{d}{dt} \left( \frac{\partial T}{\partial q^{k}} \right) = \sum_{\alpha=1}^{N} \left[ \left( \sum_{j} \frac{\partial^{2} G_{\alpha}}{\partial q^{j} \partial q^{k}} \dot{q}^{j} \right) + \frac{\partial^{2} G_{\alpha}}{\partial t \partial q^{k}} \right] \cdot m_{\alpha} x^{n}_{\alpha}
\]

\[
+ \sum_{\alpha=1}^{N} \frac{\partial G_{\alpha}}{\partial q^{k}} \cdot m_{\alpha} x^{n}_{\alpha}
\]

\[
= \sum_{\alpha=1}^{N} \left[ \left( \sum_{j} \frac{\partial^{2} G_{\alpha}}{\partial q^{j} \partial q^{k}} \dot{q}^{j} \right) + \frac{\partial^{2} G_{\alpha}}{\partial t \partial q^{k}} \right] \cdot m_{\alpha} x^{n}_{\alpha} +
\]

\[
+ \sum_{\alpha=1}^{N} \frac{\partial G_{\alpha}}{\partial q^{k}} \cdot F_{\alpha}
\]

\[
= \sum_{\alpha=1}^{N} \left[ \left( \sum_{j} \frac{\partial^{2} G_{\alpha}}{\partial q^{j} \partial q^{k}} \dot{q}^{j} + \frac{\partial^{2} G_{\alpha}}{\partial t \partial q^{k}} \right) \cdot m_{\alpha} \left( \sum_{r} \frac{\partial G_{\alpha}}{\partial q^{r}} \dot{q}^{r} + \frac{\partial G_{\alpha}}{\partial t} \right) \right] + \sum_{\alpha=1}^{N} \frac{\partial G_{\alpha}}{\partial q^{k}} \cdot F_{\alpha}
\]

(17.20)
\[ T = \sum_{j,r} \frac{1}{2} \left[ \alpha \sum_{\alpha} m_{\alpha} \left( \frac{\partial G_{\alpha}}{\partial q^j} \cdot \frac{\partial G_{\alpha}}{\partial q^r} \right) \right] \dot{q}^j \dot{q}^r + \sum_{\alpha} \sum_{j} m_{\alpha} \left( \frac{\partial G_{\alpha}}{\partial q^j} \cdot \frac{\partial G_{\alpha}}{\partial q^r \partial t} \right) \dot{q}^j + \sum_{\alpha} \sum_{j} m_{\alpha} \left( \frac{\partial G_{\alpha}}{\partial q^j} \cdot \frac{\partial^2 G_{\alpha}}{\partial q^r \partial t^2} \right) \dot{q}^j \]

\[ \frac{\partial T}{\partial q^k} = \sum_{j,r} m_{\alpha} \left( \frac{\partial^2 G_{\alpha}}{\partial q^j \partial q^r} \cdot \frac{\partial G_{\alpha}}{\partial q^k} \right) \dot{q}^j \dot{q}^r + \sum_{\alpha} \sum_{j} m_{\alpha} \left( \frac{\partial G_{\alpha}}{\partial q^j} \cdot \frac{\partial G_{\alpha}}{\partial q^k \partial t} \right) \dot{q}^j + \sum_{\alpha} m_{\alpha} \left( \frac{\partial^2 G_{\alpha}}{\partial q^j \partial t^2} \cdot \frac{\partial G_{\alpha}}{\partial q^k} \right). \]

Now upon comparing (17.23) and (17.21)

\[ \frac{d}{dt} \left( \frac{\partial T}{\partial q^k} \right) - \frac{\partial T}{\partial q^k} = \sum_{\alpha} \frac{\partial G_{\alpha}}{\partial q^k} \cdot F_{\alpha} \]

Resolve the force, \( F_{\alpha} \), into the sum of two forces, \( F_{\alpha} = F_{\alpha}^c + F_{\alpha}^e \) where \( F_{\alpha}^c \) is a force of constraint which is perpendicular to \( \frac{\partial G_{\alpha}}{\partial q^k} \) and the other force, \( F_{\alpha}^e \) which is left over is called the applied force. The applied force is allowed to have a component which is perpendicular to \( \frac{\partial G_{\alpha}}{\partial q^k} \). The only requirement of this sort is placed on \( F_{\alpha}^c \). Therefore,

\[ \frac{\partial G_{\alpha}}{\partial q^k} \cdot F_{\alpha} = \frac{\partial G_{\alpha}}{\partial q^k} \cdot F_{\alpha}^e \]

and so in the end, you obtain the following interesting equation which is equivalent to Newton’s second law.

\[ \frac{d}{dt} \left( \frac{\partial T}{\partial q^k} \right) - \frac{\partial T}{\partial q^k} = \sum_{\alpha} \frac{\partial G_{\alpha}}{\partial q^k} \cdot F_{\alpha}^e \]

(17.24)

\[ = \frac{\partial G}{\partial q^k} \cdot F^e, \]

(17.25)

where \( F^e \equiv (F^e_1, \ldots, F^e_N) \) is referred to as the total applied force.

It is particularly agreeable when the total applied force comes as the gradient of a potential function. This means there exists a scalar function of \( \mathbf{x}, \phi \) defined near \( \mathbf{G}(\mathbf{V}) \) such that

\[ F^e_{\alpha}(\mathbf{x},t) = -\nabla_\alpha \phi(\mathbf{x},t) \]
where the symbol $\nabla_\alpha$ denotes the gradient with respect to $x_\alpha$. More generally,

$$F_\alpha^a(x,t) = -\nabla_\alpha \phi(x,t) + F_\alpha^d$$

where $F_\alpha^d$ is a force which is not a force of constraint or the gradient of a given function. For example, it could be a force of friction. Then

$$F_\alpha^a(x,t) = -\nabla_\alpha \phi(x,t) + F^d$$

where

$$F^d = (F_1^d, \ldots, F_N^d)$$

Now let $T(q, \dot{q}) - \phi(G(q,t)) = L(q, \dot{q})$. Then letting $x^j$ denote the usual Cartesian coordinates of $x$,

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}^k} \right) - \frac{\partial L}{\partial x^k} = \frac{d}{dt} \left( \frac{\partial T}{\partial \dot{q}^k} \right) - \frac{\partial T}{\partial q^k} + \sum_j \frac{\partial \phi(x)}{\partial x^j} \frac{\partial x^j}{\partial q^k}$$

$$= \frac{\partial G}{\partial q^k} \left(-\nabla_\alpha \phi(x) + F^d\right) + \frac{\partial G}{\partial q^k} \nabla \phi = \frac{\partial G}{\partial q^k} \cdot F^d. \quad (17.26)$$

These are called Lagrange’s equations of motion and they are enormously significant because it is often possible to find the kinetic and potential energy in terms of variables $q^k$ which are meaningful for a particular problem. The expression, $L(q, \dot{q})$ is called the Lagrangian. This has proved part of the following theorem.

**Theorem 17.5.1** In the above context Newton’s second law implies

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{q}^k} \right) - \frac{\partial L}{\partial q^k} = \frac{\partial G}{\partial q^k} \cdot F^d. \quad (17.27)$$

In particular, if the applied force is the gradient of $-\phi$, the right side reduces to $0(17.27)$. If, in addition to this, the potential function is time independent then the total energy is conserved. That is,

$$T(q, \dot{q}) + \phi(G(q,t)) = C \quad (17.28)$$

for some constant, $C$.

**Proof:** It remains to verify the assertion about the energy. In terms of the Cartesian coordinates,

$$E = \sum_\alpha \frac{1}{2} m_\alpha \dot{x}_\alpha \cdot \dot{x}_\alpha + \phi(x,t).$$

Recall the applied force is given by $F_\alpha^a = -\nabla_\alpha \phi(x,t) + F_\alpha^d$. Differentiating with respect to time,

$$\frac{dE}{dt} = \sum_\alpha m_\alpha \ddot{x}_\alpha \cdot \dot{x}_\alpha + \sum_j \frac{\partial \phi}{\partial x^j} \ddot{x}^j + \frac{\partial \phi}{\partial t}$$

$$= \sum_\alpha F_\alpha \cdot \dot{x}_\alpha + \sum_\alpha \nabla_\alpha \phi(x,t) \cdot \dot{x}_\alpha + \frac{\partial \phi}{\partial t}$$

$$= \sum_\alpha F_\alpha^a \cdot \dot{x}_\alpha + \sum_\alpha \nabla_\alpha \phi(x,t) \cdot \dot{x}_\alpha + \frac{\partial \phi}{\partial t}$$

$$= \sum_\alpha \left(-\nabla_\alpha \phi(x,t) + F_\alpha^d\right) \cdot \dot{x}_\alpha + \sum_\alpha \nabla_\alpha \phi(x,t) \cdot \dot{x}_\alpha + \frac{\partial \phi}{\partial t}$$

$$= \sum_\alpha F_\alpha^d \cdot \dot{x}_\alpha + \frac{\partial \phi}{\partial t}.$$
Therefore, this shows (17.28) because in the case described, \( F^d = 0 \) and \( \partial \phi / \partial t = 0 \). In the case of friction, \( F^d \cdot \dot{x}_n \leq 0 \) and so in this case, if \( \phi \) is time independent, the total energy is decreasing.

**Example 17.5.2** Consider the double pendulum.

![Double Pendulum Diagram]

1. It is fairly easy to find the equations of motion in terms of the variables, \( \phi \) and \( \theta \). These variables are the \( q^k \) mentioned above. Because the two rods joining the masses have fixed length, a constraint is introduced on the motion of the two masses. It is clear the position of these masses is specified from the two variables, \( \theta \) and \( \phi \). In fact, letting the origin be located at the point at the top where the pendulum is suspended and assuming the vibration is in a plane,

\[
\mathbf{x}_1 = (l_1 \sin \theta, -l_1 \cos \theta)
\]

and

\[
\mathbf{x}_2 = (l_1 \sin \theta + l_2 \sin \phi, -l_1 \cos \theta - l_2 \cos \phi).
\]

Therefore,

\[
\dot{\mathbf{x}}_1 = \left( l_1 \dot{\theta} \cos \theta, l_1 \dot{\theta} \sin \theta \right)
\]

\[
\dot{\mathbf{x}}_2 = \left( l_1 \dot{\theta} \cos \theta + l_2 \dot{\phi} \cos \phi, l_1 \dot{\theta} \sin \theta + l_2 \dot{\phi} \sin \phi \right).
\]

It follows the kinetic energy is given by

\[
T = \frac{1}{2} m_2 \left( 2l_1 \dot{\theta} (\cos \theta) l_2 \dot{\phi} \cos \phi + l_1^2 (\dot{\theta})^2 + 2l_1 \dot{\theta} (\sin \theta) l_2 \dot{\phi} \sin \phi + l_2^2 (\dot{\phi})^2 \right) + \frac{1}{2} m_1 \left( l_1^2 (\dot{\theta})^2 \right).
\]

There are forces of constraint acting on these masses and there is the force of gravity acting on them. The force from gravity on \( m_1 \) is \(-m_1 g\) and the force from gravity on \( m_2 \) is \(-m_2 g\). Our function, \( \phi \) is just the total potential energy. Thus \( \phi (\mathbf{q}) = m_1 g y_1 + m_2 g y_2 \). It follows that \( \phi (\mathbf{q}) = m_1 g (-l_1 \cos \theta) + m_2 g (-l_1 \cos \theta - l_2 \cos \phi) \). Therefore, the Lagrangian, \( L \), is

\[
\frac{1}{2} m_2 \left( 2l_1 l_2 \dot{\theta} \dot{\phi} (\cos (\phi - \theta)) + l_1^2 (\dot{\theta})^2 + l_2^2 (\dot{\phi})^2 \right) + \frac{1}{2} m_1 \left( l_1^2 (\dot{\theta})^2 \right)
\]

\[ - [m_1 g (-l_1 \cos \theta) + m_2 g (-l_1 \cos \theta - l_2 \cos \phi)]. \]

It now becomes an easy task to find the equations of motion in terms of the two angles, \( \theta \) and \( \phi \).

\[
\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\theta}} \right) - \frac{\partial L}{\partial \theta} = 0.
\]
\[ \theta'' (m_1 + m_2) l_1^2 + m_2 l_2 l_1 \cos (\phi - \theta) \phi'' - m_2 l_1 l_2 \sin (\phi - \theta) (\phi' - \theta') \phi' \\
+ (m_1 + m_2) g l_1 \sin \theta - m_2 l_1 l_2 \phi' \sin (\phi - \theta) \\
= \theta'' (m_1 + m_2) l_1^2 + m_2 l_2 l_1 \cos (\phi - \theta) \phi'' - m_2 l_1 l_2 \sin (\phi - \theta) \phi'^2 \\
+ (m_1 + m_2) g l_1 \sin \theta = 0. \] (17.29)

To get the other equation,

\[ \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\phi}} \right) - \frac{\partial L}{\partial \phi} = \\
\frac{d}{dt} \left[ m_2 l_1 l_2 \dot{\theta} (\cos (\phi - \theta)) + m_2 l_2^2 \dot{\phi} \right] + m_2 g l_2 \sin \phi - \left( -m_2 l_1 l_2 \dot{\theta} \dot{\phi} \sin (\phi - \theta) \right) \\
= m_2 l_1 l_2 \theta'' \cos (\phi - \theta) - m_2 l_1 l_2 \theta' \sin (\phi - \theta) (\phi' - \theta') + m_2 l_2^2 \phi'' + m_2 g l_2 \sin \phi + m_2 l_1 l_2 \theta' \sin (\phi - \theta) \\
= m_2 l_1 l_2 \theta'' \cos (\phi - \theta) + m_2 l_1 l_2 (\theta')^2 \sin (\phi - \theta) + m_2 l_2^2 \phi'' + m_2 g l_2 \sin \phi = 0 \] (17.30)

Admittedly, (17.29) and (17.30) are horrific equations, but what would you expect from something as complicated as the double pendulum? They can at least be solved numerically.

The conservation of energy gives some idea what is going on. Thus

\[ \frac{1}{2} m_2 \left( 2 l_1 \dot{\theta} \dot{\phi} (\cos (\phi - \theta)) + l_1^2 (\dot{\theta})^2 + l_2^2 (\dot{\phi})^2 \right) + \frac{1}{2} m_1 \left( l_1^2 (\dot{\theta})^2 \right) \\
+ [m_1 g (-l_1 \cos \theta) + m_2 g (-l_1 \cos \theta - l_2 \cos \phi)] = C. \]

17.6 Newton’s Method

17.6.1 The Newton Raphson Method In One Dimension

The Newton Raphson method is a way to get approximations of solutions to various equations. For example, suppose you want to find \( \sqrt{2} \). The existence of \( \sqrt{2} \) is not difficult to establish by considering the continuous function, \( f(x) = x^2 - 2 \) which is negative at \( x = 0 \) and positive at \( x = 2 \). Therefore, by the intermediate value theorem, there exists \( x \in (0, 2) \) such that \( f(x) = 0 \) and this \( x \) must equal \( \sqrt{2} \). The problem consists of how to find this number, not just to prove it exists. The following picture illustrates the procedure of the Newton Raphson method.

\[ \begin{array}{c}
\blacksquare \\
\hline \\
\phi \\
\hline \\
x_1 \\
\hline \\
x_2 \\
\hline \\
\end{array} \]

In this picture, a first approximation, denoted in the picture as \( x_1 \) is chosen and then the tangent line to the curve \( y = f(x) \) at the point \( (x_1, f(x_1)) \) is obtained. The equation of this tangent line is

\[ y - f(x_1) = f'(x_1) (x - x_1). \]

Then extend this tangent line to find where it intersects the \( x \) axis. In other words, set \( y = 0 \) and solve for \( x \). This value of \( x \) is denoted by \( x_2 \). Thus

\[ x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}. \]
This second point, $x_2$ is the second approximation and the same process is done for $x_2$ that was done for $x_1$ in order to get the third approximation, $x_3$. Thus

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}.$$

Continuing this way, yields a sequence of points, $\{x_n\}$ given by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)},$$

(17.31)

which hopefully has the property that $\lim_{n \to \infty} x_n = x$ where $f(x) = 0$. You can see from the above picture that this must work out in the case of $f(x) = x^2 - 2$.

Now carry out the computations in the above case for $x_1 = 2$ and $f(x) = x^2 - 2$. From (17.31),

$$x_2 = 2 - \frac{2}{4} = 1.5.$$

Then

$$x_3 = 1.5 - \frac{(1.5)^2 - 2}{2(1.5)} \leq 1.417,$$

$$x_4 = 1.417 - \frac{(1.417)^2 - 2}{2(1.417)} = 1.414216302046577,$$

What is the true value of $\sqrt{2}$? To several decimal places this is $\sqrt{2} = 1.414213562373095$, showing that the Newton Raphson method has yielded a very good approximation after only a few iterations, even starting with an initial approximation, 2, which was not very good.

This method does not always work. For example, suppose you wanted to find the solution to $f(x) = 0$ where $f(x) = x^{1/3}$. You should check that the sequence of iterates which results does not converge. This is because, starting with $x_1$ the above procedure yields $x_2 = -2x_1$ and so as the iteration continues, the sequence oscillates between positive and negative values as its absolute value gets larger and larger. The problem is that $f'(0)$ does not exist.

However, if $f(x_0) = 0$ and $f''(x) > 0$ for $x$ near $x_0$, you can draw a picture to show that the method will yield a sequence which converges to $x_0$ provided the first approximation, $x_1$ is taken sufficiently close to $x_0$. Similarly, if $f''(x) < 0$ for $x$ near $x_0$, then the method produces a sequence which converges to $x_0$ provided $x_1$ is close enough to $x_0$.

### 17.6.2 Newton’s Method For Nonlinear Systems

The same formula yields a procedure for finding solutions to systems of functions of $n$ variables. This is particularly interesting because you can’t make any sense of things from drawing pictures. The technique of graphing and zooming which really works well for functions of one variable is no longer available.

**Procedure 17.6.1** Suppose $f$ is a $C^1$ function of $n$ variables and $f(z) = 0$. Then to find $z$, you use the same iteration which you would use in one dimension,

$$x_{k+1} = x_k - Df(x_k)^{-1} f(x_k)$$

where $x_0$ is an initial approximation chosen close to $z$.

**Example 17.6.2** Find a solution to the nonlinear system of equations,

$$f(x, y) = \begin{pmatrix} x^3 - 3xy^2 - 3x^2 + 3y^2 + 7x - 5 \\ 3x^2y - y^3 - 6xy + 7y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$
You can verify that \((x, y) = (1, 2), (1, -2)\), and \((1, 0)\) all are solutions to the above system. Suppose then that you didn’t know this.

\[
Df(x, y) = \begin{pmatrix}
3x^2 - 3y^2 - 6x + 7 & -6xy + 6y \\
6xy - 6y & 3x^2 - 3y^2 - 6x + 7
\end{pmatrix}
\]

Start with an initial guess \((x_0, y_0) = (1, 3)\). Then the next iteration is

\[
\begin{pmatrix}
1 \\
3
\end{pmatrix} - \begin{pmatrix}
-23 & 0 \\
0 & -23
\end{pmatrix}^{-1} \begin{pmatrix}
0 \\
3
\end{pmatrix} = \begin{pmatrix}
1/23 \\
3/23
\end{pmatrix}
\]

The next iteration is

\[
\begin{pmatrix}
1/23 \\
3/23
\end{pmatrix} - \begin{pmatrix}
-3.9371837 \times 10^{-2} & 0 \\
0 & -3.9371837 \times 10^{-2}
\end{pmatrix} \begin{pmatrix}
0 \\
-18.155338
\end{pmatrix} = \begin{pmatrix}
1.0 \\
2.4156258
\end{pmatrix}
\]

I will not bother to use all the decimals in 2.4156258. The next iteration is

\[
\begin{pmatrix}
1.0 \\
2.4
\end{pmatrix} - \begin{pmatrix}
-7.5301205 \times 10^{-2} & 0 \\
0 & -7.5301205 \times 10^{-2}
\end{pmatrix} \begin{pmatrix}
0 \\
-4.224
\end{pmatrix} = \begin{pmatrix}
1.0 \\
2.0819277
\end{pmatrix}
\]

Notice how the process is converging to the solution \((x, y) = (1, 2)\). If you do one more iteration, you will be really close.

The above was pretty painful because at every step the derivative had to be re-evaluated and the inverse taken. It turns out a simpler procedure will work in which you don’t have to constantly re-evaluate the inverse of the derivative.

**Procedure 17.6.3** Suppose \(f\) is a \(C^1\) function of \(n\) variables and \(f(z) = 0\). Then to find \(z\), you can use the following iteration procedure

\[
x_{k+1} = x_k - Df(x_0)^{-1} f(x_k)
\]

where \(x_0\) is an initial approximation chosen close to \(z\).

To illustrate, I will use this new procedure on the same example.

**Example 17.6.4** Find a solution to the nonlinear system of equations,

\[
f(x, y) = \begin{pmatrix}
x^3 - 3xy^2 - 3x^2 + 3y^2 + 7x - 5 \\
x^3y - y^3 - 6xy + 7y
\end{pmatrix} = \begin{pmatrix}
0 \\
0
\end{pmatrix}
\]

You can verify that \((x, y) = (1, 2), (1, -2)\), and \((1, 0)\) all are solutions to the above system. Suppose then that you didn’t know this. Take \((x_0, y_0) = (1, 3)\) as above. Then a little computation will show

\[
Df(1, 3)^{-1} = \begin{pmatrix}
-1/23 & 0 \\
0 & -1/23
\end{pmatrix}
\]
The first iteration is then

\[
0 = \begin{pmatrix} 1 - 3(2.1160873)^2 - 3 + 3(2.1160873)^2 + 7 - 5 \\ 3(2.1160873) - (2.1160873)^3 - 6(2.1160873) + 7(2.1160873) 
\end{pmatrix} = \\
\begin{pmatrix} 100 \\ 2.3478261 
\end{pmatrix}
\]

The next iteration is

\[
\begin{pmatrix} x_1 \\ y_1 
\end{pmatrix} = \begin{pmatrix} 1 \\ 0 
\end{pmatrix} - \left( \begin{pmatrix} -1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 
\end{pmatrix} \right) \begin{pmatrix} 0 \\ -15.0 
\end{pmatrix} = \\
\begin{pmatrix} 100 \\ 2.3478261 
\end{pmatrix}
\]

The next iteration is

\[
\begin{pmatrix} x_2 \\ y_2 
\end{pmatrix} = \begin{pmatrix} 1 \\ 0 
\end{pmatrix} - \left( \begin{pmatrix} -1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 
\end{pmatrix} \right) \begin{pmatrix} 0 \\ -3.5505878 
\end{pmatrix} = \\
\begin{pmatrix} 100 \\ 2.1934527 
\end{pmatrix}
\]

The next iteration is

\[
\begin{pmatrix} x_3 \\ y_3 
\end{pmatrix} = \begin{pmatrix} 1 \\ 0 
\end{pmatrix} - \left( \begin{pmatrix} -1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 
\end{pmatrix} \right) \begin{pmatrix} 0 \\ -1.779405 
\end{pmatrix} = \\
\begin{pmatrix} 100 \\ 2.1160873 
\end{pmatrix}
\]

The next iteration is

\[
\begin{pmatrix} x_4 \\ y_4 
\end{pmatrix} = \begin{pmatrix} 1 \\ 0 
\end{pmatrix} - \left( \begin{pmatrix} -1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 
\end{pmatrix} \right) \begin{pmatrix} 0 \\ -1.0111204 
\end{pmatrix} = \\
\begin{pmatrix} 100 \\ 2.0721255 
\end{pmatrix}
\]

You see it appears to be converging to a zero of the nonlinear system. It is doing so more slowly than in the case of Newton’s method but there is less trouble involved in each step of the iteration.

Of course there is a question about how to choose the initial approximation. There are methods for doing this called homotopy methods which are based on numerical methods for differential equations. The idea for these methods is to consider the problem

\[
(1 - t)(x - x_0) + tf(x) = 0.
\]

When \( t = 0 \) this reduces to \( x = x_0 \). Then when \( t = 1 \), it reduces to \( f(x) = 0 \). The equation specifies \( x \) as a function of \( t \). Differentiating with respect to \( t \), you see that \( x \) must solve the following initial value problem,

\[
-(x - x_0) + (1 - t)x' + f(x) + tDf(x)x' = 0, \quad x(0) = x_0.
\]

where \( x' \) denotes the time derivative of the vector \( x \). Initial value problems of this sort are routinely solvable using standard numerical methods. The idea is you solve it on \([0, 1]\) and your zero is \( x(1) \). Because of roundoff error, \( x(1) \) won’t be quite right so you use it as an initial guess in Newton’s method and find the zero to great accuracy.
17.7 Convergence Questions

17.7.1 A Fixed Point Theorem

The message of this section is that under reasonable conditions amounting to an assumption that \( Df(z) \) exists, Newton’s method will converge whenever you take an initial approximation sufficiently close to \( z \). This is just like the situation for the method in one dimension.

The proof of convergence rests on the following lemma which is somewhat more interesting than Newton’s method. It is a case of the contraction mapping principle important in differential and integral equations.

**Lemma 17.7.1** Suppose \( T : B(x_0, \delta) \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^p \) and it satisfies

\[
|T x - T y| \leq \frac{1}{2} |x - y| \quad \text{for all } x, y \in B(x_0, \delta). \tag{17.32}
\]

Suppose also that \( |T x_0 - x_0| < \frac{\delta}{2} \). Then \( \{T^n x_0\}_{n=1}^\infty \) converges to a point, \( x \in B(x_0, \delta) \) such that \( T x = x \). This point is called a fixed point. Furthermore, there is at most one fixed point on \( B(x_0, \delta) \).

**Proof:** From the triangle inequality, and the use of (17.32),

\[
|T^n x_0 - x_0| \leq \sum_{k=1}^{n} |T^k x_0 - T^{k-1} x_0| \\
\leq \sum_{k=1}^{n} \left( \frac{1}{2} \right)^{k-1} |T x_0 - x_0| \\
\leq 2 |T x_0 - x_0| < 2 \frac{\delta}{4} = \frac{\delta}{2} < \delta.
\]

Thus the sequence remains in the closed ball, \( B(x_0, \delta/2) \subseteq B(x_0, \delta) \). Also, by similar reasoning,

\[
|T^n x_0 - T^m x_0| \leq \sum_{k=m}^{n} |T^{k+1} x_0 - T^k x_0| \leq \sum_{k=m}^{n} \left( \frac{1}{2} \right)^k |T x_0 - x_0| \\
\leq \frac{\delta}{4} \frac{1}{2^{m-1}}.
\]

It follows, that \( \{T^n x_0\} \) is a Cauchy sequence. Therefore, it converges to a point of \( B(x_0, \delta/2) \subseteq B(x_0, \delta) \). Call this point, \( x \). Then since \( T \) is continuous, it follows \( x = \lim_{n \to \infty} T^n x_0 = T \lim_{n \to \infty} T^{n-1} x_0 = T x_0 \). If \( T x = x \) and \( T y = y \) for \( x, y \in B(x_0, \delta) \) then \( |x - y| = |T x - T y| \leq \frac{1}{2} |x - y| \) and so \( x = y \).
17.7.2 The Operator Norm

How do you measure the distance between linear transformations defined on \( \mathbb{F}^n \)? It turns out there are many ways to do this but I will give the most common one here.

**Definition 17.7.2** \( \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m) \) denotes the space of linear transformations mapping \( \mathbb{F}^n \) to \( \mathbb{F}^m \). For \( A \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m) \), the operator norm is defined by

\[
||A|| \equiv \max \{|Ax|_{\mathbb{F}^m} : |x|_{\mathbb{F}^n} \leq 1\} < \infty.
\]

**Theorem 17.7.3** Denote by \(|\cdot|\) the norm on either \( \mathbb{F}^n \) or \( \mathbb{F}^m \). Then \( \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m) \) with this operator norm is a complete normed linear space of dimension \( nm \) with

\[
||Ax|| \leq ||A|| |x|.
\]

Here Completeness means that every Cauchy sequence converges.

**Proof:** It is necessary to show the norm defined on \( \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m) \) really is a norm. This means it is necessary to verify

\[
||A|| \geq 0 \text{ and equals zero if and only if } A = 0.
\]

For \( \alpha \) a scalar,

\[
||\alpha A|| = |\alpha||A|,
\]

and for \( A, B \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m) \),

\[
||A + B|| \leq ||A|| + ||B||
\]

The first two properties are obvious but you should verify them. It remains to verify the norm is well defined and also to verify the triangle inequality above. First if \(|x| \leq 1\), and \((A_{ij})\) is the matrix of the linear transformation with respect to the usual basis vectors, then

\[
||A|| = \max \left\{ \left( \sum_i |(Ax)_i|^2 \right)^{1/2} : |x| \leq 1 \right\}
\]

\[
= \max \left\{ \left( \sum_i \sum_j A_{ij}x_j^2 \right)^{1/2} : |x| \leq 1 \right\}
\]

which is a finite number by the extreme value theorem.

It is clear that a basis for \( \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m) \) consists of linear transformations whose matrices are of the form \( E_{ij} \) where \( E_{ij} \) consists of the \( m \times n \) matrix having all zeros except for a 1 in the \( ij^{\text{th}} \) position. In effect, this considers \( \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m) \) as \( \mathbb{F}^{nm} \). Think of the \( m \times n \) matrix as a long vector folded up.

If \( x \neq 0 \),

\[
|Ax| \frac{1}{|x|} = \left| \frac{A}{|x|} x \right| \leq ||A||
\]  \hspace{1cm} (17.33)

It only remains to verify completeness. Suppose then that \( \{A_k\} \) is a Cauchy sequence in \( \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m) \). Then from (17.33) \( \{A_kx\} \) is a Cauchy sequence for each \( x \in \mathbb{F}^n \). This follows because

\[
|A_k x - A_l x| \leq ||A_k - A_l|| |x|
\]
which converges to 0 as \( k, l \to \infty \). Therefore, by completeness of \( \mathbb{F}^m \), there exists \( A_{\mathbf{x}} \), the name of the thing to which the sequence, \( \{ A_k \mathbf{x} \} \) converges such that

\[
\lim_{k \to \infty} A_k \mathbf{x} = A \mathbf{x}.
\]

Then \( A \) is linear because

\[
A(\mathbf{a} \mathbf{x} + \mathbf{b} \mathbf{y}) = \lim_{k \to \infty} A_k (\mathbf{a} \mathbf{x} + \mathbf{b} \mathbf{y}) = a \lim_{k \to \infty} A_k \mathbf{x} + b \lim_{k \to \infty} A_k \mathbf{y} = a A \mathbf{x} + b A \mathbf{y}.
\]

By the first part of this argument, \( ||A|| < \infty \) and so \( A \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m) \). This proves the theorem.

The following is an interesting exercise which is left for you.

**Proposition 17.7.4** Let \( A(\mathbf{x}) \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m) \) for each \( \mathbf{x} \in U \subseteq \mathbb{F}^p \). Then letting \( (A_{ij}(\mathbf{x})) \) denote the matrix of \( A(\mathbf{x}) \) with respect to the standard basis, it follows \( A_{ij} \) is continuous at \( \mathbf{x} \) for each \( i, j \) if and only if for all \( \varepsilon > 0 \), there exists a \( \delta > 0 \) such that if \( ||\mathbf{x} - \mathbf{y}|| < \delta \), then \( ||A(\mathbf{x}) - A(\mathbf{y})|| < \varepsilon \). That is, \( A \) is a continuous function having values in \( \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m) \) at \( \mathbf{x} \).

**Proof:** Suppose first the second condition holds. Then from the material on linear transformations,

\[
|A_{ij}(\mathbf{x}) - A_{ij}(\mathbf{y})| = |e_i \cdot (A(\mathbf{x}) - A(\mathbf{y})) e_j| \leq |e_i| ||A(\mathbf{x}) - A(\mathbf{y})|| e_j| \leq ||A(\mathbf{x}) - A(\mathbf{y})||.
\]

Therefore, the second condition implies the first.

Now suppose the first condition holds. That is each \( A_{ij} \) is continuous at \( \mathbf{x} \). Let \( ||\mathbf{v}|| \leq 1 \).

\[
||A(\mathbf{x}) - A(\mathbf{y})||(\mathbf{v}) = \left( \sum_i \sum_j |(A_{ij}(\mathbf{x}) - A_{ij}(\mathbf{y})) v_j|^2 \right)^{1/2} \tag{17.34}
\]

\[
\leq \left( \sum_i \left( \sum_j |A_{ij}(\mathbf{x}) - A_{ij}(\mathbf{y})||v_j|| \right)^2 \right)^{1/2}.
\]

By continuity of each \( A_{ij} \), there exists a \( \delta > 0 \) such that for each \( i, j \)

\[
|A_{ij}(\mathbf{x}) - A_{ij}(\mathbf{y})| < \frac{\varepsilon}{n \sqrt{m}}
\]

whenever \( ||\mathbf{x} - \mathbf{y}|| < \delta \). Then from (17.34), if \( ||\mathbf{x} - \mathbf{y}|| < \delta \),

\[
||A(\mathbf{x}) - A(\mathbf{y})||(\mathbf{v}) \leq \left( \sum_i \left( \sum_j \frac{\varepsilon}{n \sqrt{m} ||\mathbf{v}||} \right)^2 \right)^{1/2} \leq \left( \sum_i \left( \sum_j \frac{\varepsilon}{n \sqrt{m}} \right)^2 \right)^{1/2} = \varepsilon
\]
Suppose $\delta > 0$ such that if $|x - y| < \delta$, then $||Df(x) - Df(y)|| < \varepsilon$.

The proposition implies that a function is $\mathcal{C}^1$ if and only if the derivative, $Df$ exists and the function, $x \to Df(x)$ is continuous in the usual way. That is, for all $\varepsilon > 0$ there exists $\delta > 0$ such that if $|x - y| < \delta$, then $|Df(x) - Df(y)| < \varepsilon$.

This proves the proposition.

The following is a version of the mean value theorem valid for functions defined on $\mathbb{R}^n$.

**Theorem 17.7.5** Suppose $U$ is an open subset of $\mathbb{R}^p$ and $f : U \to \mathbb{R}^q$ 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

$$S \equiv \{ t \in [0, 1] : \text{for all } s \in [0, t], \}$$

$$||f(x + s (y - x)) - f(x)|| \leq (M + \varepsilon) s ||y - x||.$$  

Then $0 \in S$ and by continuity of $f$, it follows that if $t \equiv \sup S$, then $t \in S$ and if $t < 1$,

$$||f(x + t (y - x)) - f(x)|| = (M + \varepsilon) t ||y - x||.$$  \hspace{1cm} (17.35)

If $t < 1$, then there exists a sequence of positive numbers, $\{h_k\}_{k=1}^{\infty}$ converging to 0 such that

$$||f(x + (t + h_k) (y - x)) - f(x)|| > (M + \varepsilon) (t + h_k) ||y - x||$$

which implies that

$$||f(x + (t + h_k) (y - x)) - f(x + t (y - x))||$$

$$+ ||f(x + t (y - x)) - f(x)|| > (M + \varepsilon) (t + h_k) ||y - x||.$$  

By (17.35), this inequality implies

$$||f(x + (t + h_k) (y - x)) - f(x + t (y - x))|| > (M + \varepsilon) h_k ||y - x||$$

which yields upon dividing by $h_k$ and taking the limit as $h_k \to 0$,

$$||Df(x + t (y - x)) (y - x)|| \geq (M + \varepsilon) ||y - x||.$$  

Now by the definition of the norm of a linear operator,

$$M ||y - x|| \geq ||Df(x + t (y - x))|| ||y - x|| \geq ||Df(x + t (y - x)) (y - x)|| \geq (M + \varepsilon) ||y - x||,$$  

a contradiction. Therefore, $t = 1$ and so

$$||f(x + (y - x)) - f(x)|| \leq (M + \varepsilon) ||y - x||.$$  

Since $\varepsilon > 0$ is arbitrary, this proves the theorem.
17.7.3 A Method For Finding Zeros

**Theorem 17.7.6** Suppose \( f : U \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^p \) is a \( C^1 \) function and suppose \( f(z) = 0 \). Suppose also that for all \( x \) sufficiently close to \( z \), it follows that \( Df(x)^{-1} \) exists. Let \( \delta > 0 \) be small enough that for all \( x, x_0 \in B(z, 2\delta) \)

\[
\|I - Df(x_0)^{-1}Df(x)\| < \frac{1}{2}. \tag{17.36}
\]

Now pick \( x_0 \in B(z, \delta) \) also close enough to \( z \) such that

\[
\left\|Df(x_0)^{-1}\right\|f(x_0)\| \leq \frac{\delta}{4}.
\]

Define

\[ Tx \equiv x - Df(x_0)^{-1}f(x). \]

Then the sequence, \( \{T^nx_0\}_{n=1}^{\infty} \), converges to \( z \).

**Proof:** First note that \( |Tx_0 - x_0| = \left|Df(x_0)^{-1}f(x_0)\right| \leq \left\|Df(x_0)^{-1}\right\|f(x_0)\| \leq \frac{\delta}{4} \). Also on \( B(x_0, \delta) \subseteq B(z, 2\delta) \) the inequality, (17.36), the chain rule, and Theorem 17.7.5 shows that for \( x, y \in B(x_0, \delta) \),

\[
|Tx - Ty| \leq \frac{1}{2} |x - y|.
\]

This follows because \( DTx = I - Df(x_0)^{-1}f(x) \). The conclusion now follows from Lemma 17.7.1. This proves the lemma.

17.7.4 Newton’s Method

**Theorem 17.7.7** Suppose \( f : U \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^p \) is a \( C^1 \) function and suppose \( f(z) = 0 \). Suppose also that\(^2\)

\[
\left\|Df(x_2)^{-1} - Df(x_1)^{-1}\right\| \leq K|\mathbf{x}_2 - \mathbf{x}_1|. \tag{17.37}
\]

Then there exists \( \delta > 0 \) be small enough that for all \( x_1, x_2 \in B(z, 2\delta) \)

\[
|\mathbf{x}_1 - \mathbf{x}_2 - Df(x_2)^{-1}(f(x_1) - f(x_2))| \leq \frac{1}{4}|\mathbf{x}_1 - \mathbf{x}_2|, \tag{17.38}
\]

\[
|f(x_1)| < \frac{1}{4K}. \tag{17.39}
\]

Now pick \( x_0 \in B(z, \delta) \) also close enough to \( z \) such that

\[
\left\|Df(x_0)^{-1}\right\|f(x_0)\| \leq \frac{\delta}{4}.
\]

Define

\[ Tx \equiv x - Df(x)^{-1}f(x). \]

Then the sequence, \( \{T^nx_0\}_{n=1}^{\infty} \), converges to \( z \).

\(^2\)The following condition as well as the preceding can be shown to hold if you simply assume \( f \) is a \( C^2 \) function and \( Df(z)^{-1} \) exists. This requires the use of the inverse function theorem, one of the major theorems which should be studied in an advanced calculus class.
The desired result now follows from Lemma 17.7.1. This proves (17.38). (17.39) can be satisfied by taking 

because (17.37) implies \( |Df(x)|^{-1} \) is bounded for \( x \in B(z, \delta) \). Now use the assumption that \( f \) is \( C^1 \) and Proposition 17.7.4 to conclude there exists \( \delta \) small enough that \( |Df(x) - Df(z)| < \frac{1}{2} \) for all \( x \in B(z, 2\delta) \). Then let \( x_1, x_2 \in B(z, \delta) \). Define \( h(x) \equiv f(x) - f(x_2) - Df(x_2)(x-x_2) \). Then

\[
\|Dh(x)\| = \|Df(x) - Df(x_2)\| \\
\leq \|Df(x) - Df(z)\| + \|Df(z) - Df(x_2)\| \\
\leq \frac{1}{8} + \frac{1}{8} = \frac{1}{4}.
\]

It follows from Theorem 17.7.5

\[
|h(x_1) - h(x_2)| = |f(x_1) - f(x_2) - Df(x_2)(x_1 - x_2)| \\
\leq \frac{1}{4}|x_1 - x_2|.
\]

This proves (17.38). (17.39) can be satisfied by taking \( \delta \) still smaller if necessary and using \( f(z) = 0 \) and the continuity of \( f \).

Now let \( x_0 \in B(z, \delta) \) be as described. Then

\[
|Tx_0 - x_0| = \|Df(x_0)^{-1}f(x_0)\| \leq \|Df(x_0)^{-1}\| |f(x_0)| < \frac{\delta}{4}.
\]

Letting \( x_1, x_2 \in B(x_0, \delta) \subseteq B(z, 2\delta) \),

\[
|Tx_1 - Tx_2| = |x_1 - Df(x_1)^{-1}f(x_1) - (x_2 - Df(x_2)^{-1}f(x_2))| \\
\leq |x_1 - x_2 - Df(x_2)^{-1}(f(x_1) - f(x_2))| + \left|\left(Df(x_1)^{-1} - Df(x_2)^{-1}\right) f(x_1)\right| \\
\leq \frac{1}{4}|x_1 - x_2| + K |x_1 - x_2| |f(x_1)| \leq \frac{1}{2}|x_1 - x_2|.
\]

The desired result now follows from Lemma 17.7.1.

### 17.8 Exercises

1. Suppose \( f: U \to \mathbb{R}^q \) and let \( x \in U \) and \( v \) be a unit vector. Show \( D_x f(x) = Df(x) v \).

   Recall that

   \[
   D_x f(x) \equiv \lim_{h \to 0} \frac{f(x + tv) - f(x)}{t}.
   \]

2. Let \( f(x, y) = \begin{cases} xy \sin \left(\frac{1}{x}\right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases} \). Find where \( f \) is differentiable and compute the derivative at all these points.
3. Let 
\[ f(x, y) = \begin{cases} x & \text{if } |y| > |x| \\ -x & \text{if } |y| \leq |x| \end{cases} \]
Show \( f \) is continuous at \((0, 0)\) and that the partial derivatives exist at \((0, 0)\) but the function is not differentiable at \((0, 0)\).

4. Let 
\[ f(x, y, z) = \left( \frac{x^2 \sin y + z^3}{\sin (x + y) + z^3 \cos x} \right) \]
Find \( Df(1, 2, 3) \).

5. Let 
\[ f(x, y, z) = \left( \frac{x \tan y + z^3}{\cos (x + y) + z^3 \cos x} \right) \]
Find \( Df(1, 2, 3) \).

6. Let 
\[ f(x, y, z) = \left( \frac{x \sin y + z^3}{\sin (x + y) + z^3 \cos x} \right) \]
Find \( Df(x, y, z) \).

7. Let 
\[ f(x, y) = \begin{cases} (\frac{x^2 - y^2}{(x^2 + y^2)^2})^2 & \text{if } (x, y) \neq (0, 0) \\ 1 & \text{if } (x, y) = (0, 0) \end{cases} \]
Show that all directional derivatives of \( f \) exist at \((0, 0)\), and are all equal to zero but the function is not even continuous at \((0, 0)\). Therefore, it is not differentiable. Why?

8. In the example of Problem 7 show the partial derivatives exist but are not continuous.

9. A certain building is shaped like the top half of the ellipsoid, \( \frac{x^2}{900} + \frac{y^2}{900} + \frac{z^2}{400} = 1 \) determined by letting \( z \geq 0 \). Here dimensions are measured in meters. The building needs to be painted. The paint, when applied is about .005 meters thick. About how many cubic meters of paint will be needed. \textbf{Hint:} This is going to replace the numbers, 900 and 400 with slightly larger numbers when the ellipsoid is fattened slightly by the paint. The volume of the top half of the ellipsoid, \( \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} \leq 1, z \geq 0 \) is \( \frac{2}{3} abc \).

10. Show carefully that the usual one variable version of the chain rule is a special case of Theorem 17.4.2.

11. Let \( z = f(y) = (y_1^2 + \sin y_2 + \tan y_3) \) and \( y = g(x) \equiv \begin{pmatrix} x_1 + x_2 \\ x_2 - x_1 + x_2 \\ x_2^2 + x_1 + \sin x_2 \end{pmatrix} \). Find \( D(f \circ g)(x) \). Use to write \( \frac{\partial z}{\partial x_i} \) for \( i = 1, 2 \).

12. Let \( z = f(y) = (y_1^2 + \cot y_2 + \sin y_3) \) and \( y = g(x) \equiv \begin{pmatrix} x_1 + x_4 + x_3 \\ x_2^2 - x_1 + x_2 \\ x_2^2 + x_1 + \sin x_4 \end{pmatrix} \). Find \( D(f \circ g)(x) \). Use to write \( \frac{\partial z}{\partial x_i} \) for \( i = 1, 2, 3, 4 \).
Let the ideal gas law is $\frac{PV}{kT} = \text{constant}$, where $P$ is the pressure, $V$ is the volume, $T$ is the temperature, and $k$ is Boltzmann’s constant.

13. Let $z = f(y) = (y_1^2 + y_2^2 + \sin y_3 + y_4)$ and $y = g(x) \equiv \left( \begin{array}{c} x_1 + x_4 + x_3 \\ x_2 - x_1 + x_2 \\ x_3 + x_1 + \sin x_2 \\ x_4 + x_2 \end{array} \right)$. Find $D(f \circ g)(x)$. Use to write $\frac{\partial z}{\partial x_i}$ for $i = 1, 2, 3, 4$.

14. Let $z = f(y) = \left( \begin{array}{c} y_1^2 + \sin y_2 + \tan y_3 \\ y_1^2 y_2 + y_3 \\ \cos(y_1^3) + y_2^2 y_3 \end{array} \right)$ and $y = g(x) \equiv \left( \begin{array}{c} x_1 + x_4 \\ x_2 - x_1 + x_2 \\ x_3 + x_1 + \sin x_2 \end{array} \right)$. Find $D(f \circ g)(x)$. Use to write $\frac{\partial z}{\partial x_i}$ for $i = 1, 2$ and $k = 1, 2$.

15. Let $z = f(y) = \left( \begin{array}{c} y_1^2 + \sin y_2 + \tan y_3 \\ y_1^2 y_2 + y_3 \\ \cos(y_1^3) + y_2^2 y_3 \end{array} \right)$ and $y = g(x) \equiv \left( \begin{array}{c} x_1 + x_4 \\ x_2 - x_1 + x_2 \\ x_3 + x_1 + \sin x_2 \end{array} \right)$. Find $D(f \circ g)(x)$. Use to write $\frac{\partial z}{\partial x_i}$ for $i = 1, 2, 3, 4$ and $k = 1, 2, 3$.

16. Let $z = f(y) = \left( \begin{array}{c} y_1^2 + \sin y_1 + \sec y_2 + y_4 \\ y_1^2 y_2 + y_3 \\ y_1^3 y_4 + y_1 \\ y_1 + y_2 \end{array} \right)$ and $y = g(x) \equiv \left( \begin{array}{c} x_1 + x_4 \\ x_2 - x_1 + x_2 \\ x_3 + x_1 + \cos x_1 \end{array} \right)$. Find $D(f \circ g)(x)$. Use to write $\frac{\partial z}{\partial x_i}$ for $i = 1, 2, 3, 4$ and $k = 1, 2, 3$.

17. Let $f(y) = \left( \begin{array}{c} y_1^2 + \sin y_2 + \tan y_3 \\ y_1^2 y_2 + y_3 \\ \cos(y_1^3) + y_2^2 y_3 \end{array} \right)$ and $y = g(x) \equiv \left( \begin{array}{c} x_1 + x_4 \\ x_2 - x_1 + x_2 \\ x_3 + x_1 + \sin x_2 \end{array} \right)$. Find $D(f \circ g)(x)$. Use to write $\frac{\partial z}{\partial x_i}$ for $i = 1, 2, 3, 4$ and $k = 1, 2, 3$.

18. Suppose $\mathbf{r}_1(t) = (\cos t, \sin t, t), \mathbf{r}_2(t) = (t, 2t, 1)$, and $\mathbf{r}_3(t) = (1, t, 1)$. Find the rate of change with respect to $t$ of the volume of the parallelepiped determined by these three vectors when $t = 1$.

19. A trash compactor is compacting a rectangular block of trash. The width is changing at the rate of $-1$ inches per second, the length is changing at the rate of $-2$ inches per second and the height is changing at the rate of $-3$ inches per second. How fast is the volume changing when the length is 20, the height is 10, and the width is 10.

20. A trash compactor is compacting a rectangular block of trash. The width is changing at the rate of $-2$ inches per second, the length is changing at the rate of $-1$ inches per second and the height is changing at the rate of $-4$ inches per second. How fast is the surface area changing when the length is 20, the height is 10, and the width is 10.

21. The ideal gas law is $PV = kT$ where $k$ is a constant which depends on the number of moles and on the gas being considered. If $V$ is changing at the rate of 2 cubic cm. per second and $T$ is changing at the rate of 3 degrees Kelvin per second, how fast is the pressure changing when $T = 300$ and $V$ equals 400 cubic cm.?

22. Let $S$ denote a level surface of the form $f(x_1, x_2, x_3) = C$. Suppose now that $\mathbf{r}(t)$ is a space curve which lies in this level surface. Thus $f(\mathbf{r}_1(t), \mathbf{r}_2(t), \mathbf{r}_3(t))$. Show using the chain rule that $Df(\mathbf{r}_1(t), \mathbf{r}_2(t), \mathbf{r}_3(t))(\mathbf{r}_1'(t), \mathbf{r}_2'(t), \mathbf{r}_3'(t)) = 0$. Note that $Df(x_1, x_2, x_3) = (f_{x_1}, f_{x_2}, f_{x_3})$. This is denoted by $\nabla f(x_1, x_2, x_3) = (f_{x_1}, f_{x_2}, f_{x_3})$. This 3 × 1 matrix or column vector is called the gradient vector. Argue that

$$\nabla f(\mathbf{r}_1(t), \mathbf{r}_2(t), \mathbf{r}_3(t)) \cdot (\mathbf{r}_1'(t), \mathbf{r}_2'(t), \mathbf{r}_3'(t)) = 0.$$
23. Suppose $f$ is a $C^1$ function which maps $U$, an open subset of $\mathbb{R}^n$ one to one and onto $V$, an open set in $\mathbb{R}^m$ such that the inverse map, $f^{-1}$ is also $C^1$. What must be true of $m$ and $n$? Why? \textbf{Hint:} Consider Example 17.4.5 on Page 312.
The Gradient

18.0.1 Outcomes

1. Interpret the gradient of a function as a normal to a level curve or a level surface.
2. Find the normal line and tangent plane to a smooth surface at a given point.
3. Find the angles between curves and surfaces.

Here we review the concept of the gradient. This has already been considered in the special case of a $C^1$ function. However, you do not need so much to talk of the gradient.

18.1 Fundamental Properties

Let $f : U \to \mathbb{R}$ where $U$ is an open subset of $\mathbb{R}^n$ and suppose $f$ is differentiable on $U$. Thus if $x \in U$,

$$f(x + v) = f(x) + \sum_{j=1}^{n} \frac{\partial f(x)}{\partial x_i} v_j + o(t).$$  \hspace{1cm} (18.1)

Recall Proposition 16.3.6, a more general version of which is stated here for convenience. It is more general because we only assume $f$ is differentiable, not $C^1$.

**Proposition 18.1.1** If $f$ is differentiable at $x$ and for $v$ a unit vector,

$$D_v f(x) = \nabla f(x) \cdot v.$$

**Proof:**

$$\frac{f(x + tv) - f(x)}{t} = \frac{1}{t} \left( f(x) + \sum_{j=1}^{n} \frac{\partial f(x)}{\partial x_i} t v_i + o(tv) - f(x) \right)$$

$$= \frac{1}{t} \left( \sum_{j=1}^{n} \frac{\partial f(x)}{\partial x_i} t v_i + o(tv) \right)$$

$$= \sum_{j=1}^{n} \frac{\partial f(x)}{\partial x_i} v_i + \frac{o(tv)}{t}.$$

Now $\lim_{t \to 0} \frac{o(tv)}{t} = 0$ and so

$$D_v f(x) = \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t} = \sum_{j=1}^{n} \frac{\partial f(x)}{\partial x_i} v_i = \nabla f(x) \cdot v.$$
as claimed.

**Definition 18.1.2** When $f$ is differentiable, define $\nabla f(x) \equiv \left( \frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right)^T$ just as was done in the special case where $f$ is $C^1$. As before, this vector is called the gradient vector.

This defines the gradient for a differentiable scalar valued function. There are ways to define the gradient for vector valued functions but this will not be attempted in this book.

It follows immediately from (18.1) that

$$f(x + \mathbf{v}) = f(x) + \nabla f(x) \cdot \mathbf{v} + o(\mathbf{v}) \quad (18.2)$$

An important aspect of the gradient is its relation with the directional derivative. From (18.2), for $\mathbf{v}$ a unit vector,

$$\frac{f(x + t\mathbf{v}) - f(x)}{t} = \nabla f(x) \cdot \mathbf{v} + \frac{o(t)}{t}$$

Therefore, taking $t \to 0$,

$$D_\mathbf{v}f(x) = \nabla f(x) \cdot \mathbf{v}. \quad (18.3)$$

**Example 18.1.3** Let $f(x, y, z) = x^2 + \sin(xy) + z$. Find $D_\mathbf{v}f(1, 0, 1)$ where $\mathbf{v} = \left( \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right)$.

Note this vector which is given is already a unit vector. Therefore, from the above, it is only necessary to find $\nabla f(1, 0, 1)$ and take the dot product. $\nabla f(x, y, z) = (2x, x \cos(xy), 1)$. Therefore, $\nabla f(1, 0, 1) = (2, 1, 1)$. Therefore, the directional derivative is $(2, 1, 1) \cdot \left( \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right) = \frac{4}{\sqrt{3}}$.

Because of (18.3) it is easy to find the largest possible directional derivative and the smallest possible directional derivative. That which follows is a more algebraic treatment of an earlier result with the trigonometry removed.

**Proposition 18.1.4** Let $f : U \to \mathbb{R}$ be a differentiable function and let $x \in U$. Then

max \{ $D_\mathbf{v}f(x) : |\mathbf{v}| = 1$ \} = $|\nabla f(x)|$ \quad (18.4)

and

min \{ $D_\mathbf{v}f(x) : |\mathbf{v}| = 1$ \} = $-|\nabla f(x)|$. \quad (18.5)

Furthermore, the maximum in (18.4) occurs when $\mathbf{v} = \nabla f(x) / |\nabla f(x)|$ and the minimum in (18.5) occurs when $\mathbf{v} = -\nabla f(x) / |\nabla f(x)|$.

**Proof:** From (18.3) and the Cauchy Schwarz inequality,

$$|D_\mathbf{v}f(x)| \leq |\nabla f(x)|$$

and so for any choice of $\mathbf{v}$ with $|\mathbf{v}| = 1$,

$$-|\nabla f(x)| \leq D_\mathbf{v}f(x) \leq |\nabla f(x)|.$$

The proposition is proved by noting that if $\mathbf{v} = -\nabla f(x) / |\nabla f(x)|$, then

$$D_\mathbf{v}f(x) = \nabla f(x) \cdot (-\nabla f(x) / |\nabla f(x)|)$$

$$= -|\nabla f(x)|^2 / |\nabla f(x)| = -|\nabla f(x)|$$
while if \( \mathbf{v} = \nabla f(x) / |\nabla f(x)| \), then
\[
D_x f(x) = \frac{\nabla f(x) \cdot (\nabla f(x) / |\nabla f(x)|)}{|\nabla f(x)|} = \frac{|\nabla f(x)|^2}{|\nabla f(x)|} = |\nabla f(x)|.
\]

The conclusion of the above proposition is important in many physical models. For example, consider some material which is at various temperatures depending on location. Because it has cool places and hot places, it is expected that the heat will flow from the hot places to the cool places. Consider a small surface having a unit normal, \( \mathbf{n} \). Thus \( \mathbf{n} \) is a normal to this surface and has unit length. If it is desired to find the rate in calories per second at which heat crosses this little surface in the direction of \( \mathbf{n} \), it is defined as \( \mathbf{J} \cdot \mathbf{n} A \) where \( A \) is the area of the surface and \( \mathbf{J} \) is called the heat flux. It is reasonable to suppose the rate at which heat flows across this surface will be largest when \( \mathbf{n} \) is in the direction of greatest rate of decrease of the temperature. In other words, heat flows most readily in the direction which involves the maximum rate of decrease in temperature. This expectation will be realized by taking \( \mathbf{J} = -K \nabla c \) where \( K \) is a positive scalar function which can depend on a variety of things. The above relation between the heat flux and \( \nabla c \) is usually called the Fourier heat conduction law and the constant, \( K \), is known as the coefficient of thermal conductivity. It is a material property, different for iron than for aluminum. In most applications, \( K \) is considered to be a constant but this is wrong. Experiments show this scalar should depend on temperature. Nevertheless, things get very difficult if this dependence is allowed. The constant can depend on position in the material or even on time.

An identical relationship is usually postulated for the flow of a diffusing species. In this problem, something like a pollutant diffuses. It may be an insecticide in ground water for example. Like heat, it tries to move from areas of high concentration toward areas of low concentration. In this case \( \mathbf{J} = -K \nabla c \) where \( c \) is the concentration of the diffusing species. When applied to diffusion, this relationship is known as Fick’s law. Mathematically, it is indistinguishable from the problem of heat flow.

Note the importance of the gradient in formulating these models.

### 18.2 Tangent Planes

The gradient has fundamental geometric significance illustrated by the following picture.

In this picture, the surface is a piece of a level surface of a function of three variables, \( f(x, y, z) \). Thus the surface is defined by \( f(x, y, z) = c \) or more completely as \( \{(x, y, z) : f(x, y, z) = c\} \). For example, if \( f(x, y, z) = x^2 + y^2 + z^2 \), this would be a piece of a sphere. There are two smooth curves in this picture which lie in the surface having parameterizations, \( x_1(t) = (x_1(t), y_1(t), z_1(t)) \) and \( x_2(s) = (x_2(s), y_2(s), z_2(s)) \) which intersect at the point, \( (x_0, y_0, z_0) \) on this surface. This intersection occurs when \( t = t_0 \) and \( s = s_0 \). Since the points, \( x_1(t) \) for \( t \) in an interval lie in the level surface, it follows
\[ f(x_t(t), y_t(t), z_t(t)) = c \]

for all \( t \) in some interval. Therefore, taking the derivative of both sides and using the chain rule on the left,

\[ \frac{\partial f}{\partial x} (x_t(t), y_t(t), z_t(t)) \frac{dx}{dt} + \frac{\partial f}{\partial y} (x_t(t), y_t(t), z_t(t)) \frac{dy}{dt} + \frac{\partial f}{\partial z} (x_t(t), y_t(t), z_t(t)) \frac{dz}{dt} = 0. \]

In terms of the gradient, this merely states

\[ \nabla f (x_t(t), y_t(t), z_t(t)) \cdot \frac{dx}{dt} = 0. \]

Similarly,

\[ \nabla f (x_2(s), y_2(s), z_2(s)) \cdot \frac{dx}{dt} = 0. \]

Letting \( s = s_0 \) and \( t = t_0 \), it follows

\[ \nabla f (x_0, y_0, z_0) \cdot \frac{dx}{dt} (t_0) = 0, \quad \nabla f (x_0, y_0, z_0) \cdot \frac{dx}{dt} (s_0) = 0. \]

It follows \( \nabla f (x_0, y_0, z_0) \) is perpendicular to both the direction vectors of the two indicated curves shown. Surely if things are as they should be, these two direction vectors would determine a plane which deserves to be called the tangent plane to the level surface of \( f \) at the point \((x_0, y_0, z_0)\) and that \( \nabla f (x_0, y_0, z_0) \) is perpendicular to this tangent plane at the point, \((x_0, y_0, z_0)\).

**Example 18.2.1** Find the equation of the tangent plane to the level surface, \( f(x, y, z) = 6 \) of the function, \( f(x, y, z) = x^2 + 2y^2 + 3z^2 \) at the point \((1, 1, 1)\).

First note that \((1, 1, 1)\) is a point on this level surface. To find the desired plane it suffices to find the normal vector to the proposed plane. But \( \nabla f (x, y, z) = (2x, 4y, 6z) \) and so \( \nabla f (1, 1, 1) = (2, 4, 6) \). Therefore, from this problem, the equation of the plane is

\[ (2, 4, 6) \cdot (x - 1, y - 1, z - 1) = 0 \]

or in other words,

\[ 2x - 12 + 4y + 6z = 0. \]

**Example 18.2.2** The point, \((\sqrt{3}, 1, 4)\) is on both the surfaces, \( z = x^2 + y^2 \) and \( z = 8 - (x^2 + y^2) \). Find the cosine of the angle between the two tangent planes at this point.

Recall this is the same as the angle between two normal vectors. Of course there is some ambiguity here because if \( n \) is a normal vector, then so is \(-n\) and replacing \( n \) with \(-n\) in the formula for the cosine of the angle will change the sign. We agree to look for the acute angle and its cosine rather than the optuse angle. The normals are \((2\sqrt{3}, 2, -1)\) and \((2\sqrt{3}, 2, 1)\). Therefore, the cosine of the angle desired is

\[ \frac{(2\sqrt{3})^2 + 4 - 1}{17} = \frac{15}{17} \]

consequence of the implicit function theorem, one of the greatest theorems in all mathematics and a topic for an advanced calculus class.
Example 18.2.3 The point, \((1, \sqrt{3}, 4)\) is on the surface, \(z = x^2 + y^2\). Find the line perpendicular to the surface at this point.

All we need is the direction vector of this line. The surface is the level surface, \(x^2 + y^2 - z = 0\). The normal to this surface is given by the gradient at this point. Thus the line desired is

\[
\left(1, \sqrt{3}, 4\right) + t \left(2, 2\sqrt{3}, -1\right).
\]

18.3 Exercises

1. Find the gradients of \(f =
   \begin{align*}
   & (a) \ x^2 y + z^3 \text{ at } (1, 1, 2) \\
   & (b) \ z \sin(x^2 y) + 2^{x+y} \text{ at } (1, 1, 0) \\
   & (c) \ u \ln(x + y + z^2 + w) \text{ at } (x, y, z, w, u) = (1, 1, 1, 2)
   \end{align*}
   
2. Find the directional derivatives of \(f\) at the indicated point in the direction, \((\frac{1}{2}, \frac{1}{2}, \frac{1}{\sqrt{2}})\).
   \begin{align*}
   & (a) \ x^2 y + z^3 \text{ at } (1, 1, 1) \\
   & (b) \ z \sin(x^2 y) + 2^{x+y} \text{ at } (1, 1, 2) \\
   & (c) \ xy + z^2 + w \text{ at } (1, 2, 3)
   \end{align*}

3. Find the tangent plane to the indicated level surface at the indicated point.
   \begin{align*}
   & (a) \ x^2 y + z^3 = 2 \text{ at } (1, 1, 1) \\
   & (b) \ z \sin(x^2 y) + 2^{x+y} = 2 \sin 1 + 4 \text{ at } (1, 1, 2) \\
   & (c) \ \cos(x) + z \sin(x + y) = 1 \text{ at } (-\pi, \frac{3\pi}{2}, 2)
   \end{align*}

4. Explain why the displacement vector of an object moving in \(\mathbb{R}^3\) is always perpendicular to the velocity vector if the object is always at a fixed distance from a given point.

5. The point \((1, 1, \sqrt{2})\) is a point on the level surface, \(x^2 + y^2 + z^2 = 4\). Find the line perpendicular to the surface at this point.

6. The point \((1, 1, \sqrt{2})\) is a point on the level surface, \(x^2 + y^2 + z^2 = 4\) and the level surface, \(y^2 + 2z^2 = 5\). Find the angle between the two tangent planes at this point.

7. The level surfaces \(x^2 + y^2 + z^2 = 4\) and \(z + x^2 + y^2 = 4\) have the point \((\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 1)\) in the curve formed by the intersection of these surfaces. Find a direction vector for this curve at this point. **Hint:** Recall the gradients of the two surfaces are perpendicular to the corresponding surfaces at this point. A direction vector for the desired curve should be perpendicular to both of these gradients.

8. In a slightly more general setting, suppose \(f_1 (x, y, z) = 0\) and \(f_2 (x, y, z) = 0\) are two level surfaces which intersect in a curve which has parameterization, \((x (t), y (t), z (t))\). Find a differential equation for this curve.
Optimization

19.0.1 Outcomes

1. Define what is meant by a local extreme point.
2. Find candidates for local extrema using the gradient.
3. Find the local extreme values and saddle points of a $C^2$ function.
4. Use the second derivative test to identify the nature of a singular point.
5. Find the extreme values of a function defined on a closed and bounded region.
6. Solve word problems involving maximum and minimum values.
7. Use the method of Lagrange to determine the extreme values of a function subject to a constraint.
8. Solve word problems using the method of Lagrange multipliers.

Suppose $f : D(f) \to \mathbb{R}$ where $D(f) \subseteq \mathbb{R}^n$.

19.1 Local Extrema

Definition 19.1.1 A point $x \in D(f)$ is called a local minimum if $f(x) \leq f(y)$ for all $y \in D(f)$ sufficiently close to $x$. A point $x \in D(f)$ is called a local maximum if $f(x) \geq f(y)$ for all $y \in D(f)$ sufficiently close to $x$. A local extremum is a point of $D(f)$ which is either a local minimum or a local maximum. The plural for extremum is extrema.

Procedure 19.1.2 To find candidates for local extrema which are interior points of $D(f)$ where $f$ is a $C^1$ function, you simply identify those points where $\nabla f$ equals the zero vector. To justify this, note that the graph of $f$ is the level surface

$$F(x,z) \equiv f(x) - z = 0$$

and the local extrema at such interior points must have horizontal tangent planes. Therefore, a normal vector at such points must be a multiple of $(0, \cdots, 0, 1)$. Thus $\nabla F$ at such points must be a multiple of this vector. That is, if $x$ is such a point,

$$k(0, \cdots, 0, 1) = (f_{x1}(x), \cdots, f_{xn}(x), -1).$$

Thus $\nabla f(x) = 0$. 

341
Definition 19.1.3 A singular point for \( f \) is a point \( x \) where \( \nabla f(x) = 0 \).

Example 19.1.4 Find the local extrema for the function, \( f(x, y) \equiv xy - x - y \) for \( x, y > 0 \).

Note that here \( D(f) \) is an open set and so every point is an interior point. Where is the gradient equal to zero?

\[
\begin{align*}
\frac{\partial f}{\partial x} &= y - 1 = 0, \\
\frac{\partial f}{\partial y} &= x - 1 = 0
\end{align*}
\]

and so there is exactly one candidate for a local extrema, \((1, 1)\).

Example 19.1.5 Find the volume of the smallest tetrahedron made up of the coordinate planes in the first octant and a plane which is tangent to the sphere \( x^2 + y^2 + z^2 = 4 \).

The normal to the sphere at a point, \((x_0, y_0, z_0)\) on a point of the sphere is \((x_0, y_0, \sqrt{4 - x_0^2 - y_0^2})\) and so the equation of the tangent plane at this point is

\[
x_0 (x - x_0) + y_0 (y - y_0) + \sqrt{4 - x_0^2 - y_0^2} \left( z - \sqrt{4 - x_0^2 - y_0^2} \right) = 0
\]

When \( x = y = 0 \),

\[
z = \frac{4}{\sqrt{4 - x_0^2 - y_0^2}}
\]

When \( z = 0 = y \),

\[
x = \frac{4}{x_0},
\]

and when \( z = x = 0 \),

\[
y = \frac{4}{y_0}.
\]

Therefore, the function to minimize is

\[
f(x, y) = \frac{1}{6} \frac{64}{xy\sqrt{(4 - x^2 - y^2)}}
\]

This is because in beginning calculus it was shown that the volume of a pyramid is \(1/3\) the area of the base times the height. Therefore, you simply need to find the gradient of this and set it equal to zero. Thus upon taking the partial derivatives, you need to have

\[
-4 + 2x^2 + y^2 \frac{\sqrt{(4 - x^2 - y^2)}}{x^2y(-4 + x^2 + y^2)} = 0,
\]

and

\[
-4 + 2x^2 + 2y^2 \frac{\sqrt{(4 - x^2 - y^2)}}{xy^2(-4 + x^2 + y^2)} = 0.
\]

Therefore, \( x^2 + 2y^2 = 4 \) and \( 2x^2 + y^2 = 4 \). Thus \( x = y \) and so \( x = y = \frac{2}{\sqrt{3}} \). It follows from the equation for \( z \) that \( z = \frac{2}{\sqrt{3}} \) also.

Example 19.1.6 An open box is to contain 32 cubic feet. Find the dimensions which will result in the least surface area.
Let the height of the box be \( z \) and the length and width be \( x \) and \( y \) respectively. Then \( xyz = 32 \) and so \( z = \frac{32}{xy} \). The total area is \( xy + 2xz + 2yz \) and so in terms of the two variables, \( x \) and \( y \), the area is
\[
A = xy + \frac{64}{y} + \frac{64}{x}.
\]
To find best dimensions you note these must result in a local minimum.
\[
A_x = \frac{yx^2 - 64}{x^2} = 0, \quad A_y = \frac{xy^2 - 64}{y^2}.
\]
Therefore, \( yx^2 - 64 = 0 \) and \( xy^2 - 64 = 0 \) so \( xy^2 = yx^2 \). For sure the answer excludes the case where any of the variables equals zero. Therefore, \( x = y \) and so \( x = 4 = y \). Then \( z = 2 \) from the requirement that \( xyz = 32 \).

19.2 The Second Derivative Test

There is a version of the second derivative test in the case that the function and its first and second partial derivatives are all continuous. A discussion of its proof is given in Section 19.3.

**Theorem 19.2.1** Let \( f : U \to \mathbb{R} \) for \( U \) an open set in \( \mathbb{R}^n \) and let \( f \) be a \( C^2 \) function and suppose that at some \( x \in U, \nabla f(x) = 0 \). Also let \( \mu \) and \( \lambda \) be respectively, the largest and smallest eigenvalues of the matrix, \( H \). If \( \lambda > 0 \) then \( f \) has a local minimum at \( x \). If \( \mu < 0 \) then \( f \) has a local maximum at \( x \). If either \( \lambda \) or \( \mu \) equals zero, the test fails. If \( \lambda < 0 \) and \( \mu > 0 \) there exists a direction in which when \( f \) is evaluated on the line through the critical point having this direction, the resulting function of one variable has a local minimum and there exists a direction in which when \( f \) is evaluated on the line through the critical point having this direction, the resulting function of one variable has a local maximum. This last case is called a saddle point.

**Example 19.2.2** Let \( f(x, y) = 2x^4 - 4x^3 + 14x^2 + 12yx^2 - 12yx - 12x + 2y^2 + 4y + 2 \). Find the critical points and determine whether they are local minima, local maxima, or saddle points.

\[
f_x(x, y) = 8x^3 - 12x^2 + 28x + 24yx - 12 - 12y = 0 \quad \text{and} \quad f_y(x, y) = 12x^2 - 12x + 4y + 4.
\]
The points at which both \( f_x \) and \( f_y \) equal zero are \((\frac{1}{2}, -\frac{1}{4})\), \((0, -1)\), and \((1, -1)\).

The Hessian matrix is
\[
\begin{pmatrix}
24x^2 + 28 + 24y - 24x & 24x - 12 \\
24x - 12 & 4
\end{pmatrix}
\]
and the thing to determine is the sign of its eigenvalues evaluated at the critical points.

First consider the point \((\frac{1}{2}, -\frac{1}{4})\). This matrix is \(\begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix}\) and its eigenvalues are 16, 4 showing that this is a local minimum.

Next consider \((0, -1)\) at this point the Hessian matrix is \(\begin{pmatrix} 4 & -12 \\ -12 & 4 \end{pmatrix}\) and the eigenvalues are 16, -8. Therefore, this point is a saddle point.

Finally consider the point \((1, -1)\). At this point the Hessian is \(\begin{pmatrix} 4 & 12 \\ 12 & 4 \end{pmatrix}\) and the eigenvalues are 16, -8 so this point is also a saddle point.

The geometric significance of a saddle point was explained above. In one direction it looks like a local minimum while in another it looks like a local maximum. In fact, they
do look like a saddle. Here is a picture of the graph of the above function near the saddle point, \((0, -1)\).

\[
\begin{align*}
&f(x, y) = \arctan\left(2x^4 - 4x^3 + 14x^2 - 12xy - 12x + 2y^2 + 4y + 2\right),
\end{align*}
\]

Since \(\arctan\) is a strictly increasing function, it preserves all the information about whether the given function is increasing or decreasing in certain directions. Below is a graph of this function which illustrates the behavior near the point \((1, -1)\).

Or course sometimes the second derivative test is inadequate to determine what is going on. This should be no surprise since this was the case even for a function of one variable. For a function of two variables, a nice example is the Monkey saddle.

\textbf{Example 19.2.3} Suppose \(f(x, y) = \arctan\left(6xy^2 - 2x^3 - 3y^4\right)\). Show \((0, 0)\) is a critical point for which the second derivative test gives no information.

Before doing anything it might be interesting to look at the graph of this function of two variables plotted using Maple.
This picture should indicate why this is called a monkey saddle. It is because the monkey can sit in the saddle and have a place for his tail. Now to see \((0, 0)\) is a critical point, note that
\[
\frac{\partial}{\partial x} \left( \arctan \left( g(x, y) \right) \right) = \frac{1}{1 + g(x, y)^2} g_x(x, y)
\]
and that a similar formula holds for the partial derivative with respect to \(y\). Therefore, it suffices to verify that for
\[
g(x, y) = 6xy^2 - 2x^3 - 3y^4
\]
\(g_x(0, 0) = g_y(0, 0) = 0\).

This reduces to
\[
h(t) = f(t, t) = \arctan(4t^3 - t^4),
\]
which is strictly increasing near \(t = 0\). Therefore, the Hessian matrix is the zero matrix and clearly has only the zero eigenvalue. Therefore, the second derivative test is totally useless at this point.

However, suppose you took \(x = t\) and \(y = t\) and evaluated this function on this line. This reduces to \(h(t) = f(t, t) = \arctan(4t^3 - t^4)\), which is strictly increasing near \(t = 0\). This shows the critical point, \((0, 0)\) of \(f\) is neither a local max. nor a local min. Next let \(x = 0\) and \(y = t\). Then \(p(t) \equiv f(0, t) = -3t^4\). Therefore, along the line, \((0, t), f\) has a local maximum at \((0, 0)\).

19.3 Proof Of The Second Derivative Test

**Definition 19.3.1** The matrix, \(\left( \frac{\partial^2 f}{\partial x_i \partial x_j} \right)(x)\) is called the Hessian matrix, denoted by \(H(x)\).

Now recall the Taylor formula with the Lagrange form of the remainder. See any good non reformed calculus book for a proof of this theorem. Ellis and Gulleck has a good proof. It is stated only for the interval on which it will be used.

**Theorem 19.3.2** Let \(h : (-\delta, 1 + \delta) \rightarrow \mathbb{R}\) have \(m + 1\) derivatives. Then there exists \(t \in (0, 1)\) such that
\[
h(1) = h(0) + \sum_{k=1}^{m} \frac{h^{(k)}(0)}{k!} + \frac{h^{(m+1)}(t)}{(m+1)!}.
\]
Suppose and let so \( H \) and by continuity of the second derivatives, these mixed second derivatives are equal and because of the continuity of the entries of \( H \), where the last term satisfies

\[
\frac{\partial^2 f}{\partial x_i \partial x_j} (x + tv) v_i v_j.
\]

Thus

\[
h''(t) = v^T H (x + tv) v.
\]

From Theorem 19.3.2 there exists \( t \in (0, 1) \) such that

\[
f(x + v) = f(x) + \frac{\partial f}{\partial x_i}(x) v_i + \frac{1}{2} v^T H(x) v + \frac{1}{2} (v^T (H(x + tv) - H(x)) v
\]

By the continuity of the second partial derivative

\[
f(x + v) = f(x) + \nabla f(x) \cdot v + \frac{1}{2} v^T H(x) v + \frac{1}{2} (v^T (H(x + tv) - H(x)) v
\]

where the last term satisfies

\[
\lim_{|v| \to 0} \frac{1}{|v|^2} \left( \frac{1}{2} (v^T (H(x + tv) - H(x)) v \right) = 0
\]

because of the continuity of the entries of \( H(x) \).

**Theorem 19.3.3** Suppose \( x \) is a critical point for \( f \). That is, suppose \( \frac{\partial f}{\partial x_i}(x) = 0 \) for each \( i \). Then if \( H(x) \) has all positive eigenvalues, \( x \) is a local minimum. If \( H(x) \) 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 \( f \) has a local maximum at \( x \).

**Proof:** Since \( \nabla f(x) = 0 \), formula (19.1) implies

\[
f(x + v) = f(x) + \frac{1}{2} v^T H(x) v + \frac{1}{2} v^T (H(x + tv) - H(x)) v
\]

and by continuity of the second derivatives, these mixed second derivatives are equal and so \( H(x) \) is a symmetric matrix. Thus, by Corollary 10.4.8 on Page 196 \( H(x) \) has all real eigenvalues. Suppose first that \( H(x) \) has all positive eigenvalues and that all are larger than \( \delta^2 > 0 \). Then by this corollary, \( H(x) \) has an orthonormal basis of eigenvectors, \( \{v_i\}_{i=1}^n \) and so if \( u \) is an arbitrary vector, there exist scalars, \( u_i \) such that \( u = \sum_{j=1}^n u_j v_j \). Taking the dot product of both sides with \( v_j \), it follows \( u_j = u \cdot v_j \). Thus

\[
\begin{align*}
u^T H(x) u &= \left( \sum_{k=1}^n u_k v_k^T \right) H(x) \left( \sum_{j=1}^n u_j v_j \right) \\
&= \sum_{k,j} u_k v_k^T H(x) v_j u_j
\end{align*}
\]
\[ \sum_{j=1}^{n} u_j^2 \lambda_j \geq \delta^2 \sum_{j=1}^{n} u_j^2 = \delta^2 |u|^2. \]

From (19.3) and (19.2), 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 (19.3), replacing \( v \) with \( sv \) and letting \( t \) depend on \( s \),
\[
f(x + sv) = f(x) + \frac{1}{2} s^2 v^T H(x) v + \frac{1}{2} s^2 (v^T (H(x + tsv) - H(x)) v)
\]
which implies
\[
f(x + sv) = f(x) + \frac{1}{2} s^2 \lambda^2 |v|^2 + \frac{1}{2} s^2 (v^T (H(x + tsv) - H(x)) v)
\]
\[
\geq f(x) + \frac{1}{4} s^2 \lambda^2 |v|^2
\]
whenever \( s \) 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.

**19.4 Exercises**

1. Use the second derivative test on the critical points \((1, 1)\), and \((1, -1)\) for Example 19.2.3.
2. If \( H = H^T \) and \( Hx = \lambda x \) while \( Hx = \mu x \) for \( \lambda \neq \mu \), show \( x \cdot y = 0 \).
3. Show the points \( \left( \frac{1}{2}, -\frac{21}{4} \right), (0, -4), \) and \((1, -4)\) are critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.
\[ f(x, y) = -x^4 + 2x^3 + 39x^2 + 10yx^2 - 10yx - 40x - y^2 - 8y - 16. \]
Answer:
The Hessian matrix is
\[
\begin{pmatrix}
-12x^2 + 78 + 20y + 12x & 20x - 10 \\
20x - 10 & -2
\end{pmatrix}
\]
The eigenvalues must be checked at the critical points. First consider the point \( \left( \frac{1}{2}, -\frac{21}{4} \right) \). At this point, the Hessian is
\[
\begin{pmatrix}
-24 & 0 \\
0 & -2
\end{pmatrix}
\]
and its eigenvalues are \(-24, -2\), both negative. Therefore, the function has a local maximum at this point.
Next consider \((0, -4)\). At this point the Hessian matrix is
\[
\begin{pmatrix}
  -2 & -10 \\
  -10 & -2
\end{pmatrix}
\]
and the eigenvalues are \(8, -12\) so the function has a saddle point.

Finally consider the point \((1, -4)\). The Hessian equals
\[
\begin{pmatrix}
  -2 & 10 \\
  10 & -2
\end{pmatrix}
\]
having eigenvalues: \(8, -12\) and so there is a saddle point here.

4. Show the points \(\left(\frac{1}{7}, -\frac{53}{72}\right), (0, -4)\), and \((1, -4)\) are critical points of the following function of three variables and classify them according to whether they are local minimums, local maximums or saddle points.

\[
f(x, y) = -3x^4 + 6x^3 + 37x^2 + 10yx^2 - 10yx - 40x - 3y^2 - 24y - 48.
\]
Answer:
The Hessian matrix is
\[
\begin{pmatrix}
  -36x^2 + 74 + 20y + 36x & 20x - 10 \\
  20x - 10 & -6
\end{pmatrix}
\]
Check its eigenvalues at the critical points. First consider the point \(\left(\frac{1}{7}, -\frac{53}{72}\right)\). At this point the Hessian is
\[
\begin{pmatrix}
  -\frac{16}{7} & 0 \\
  0 & -6
\end{pmatrix}
\]
and its eigenvalues are \(-\frac{16}{7}, -6\) so there is a local maximum at this point. The same analysis shows there are saddle points at the other two critical points.

5. Show the points \(\left(\frac{1}{7}, \frac{37}{20}\right), (0, 2)\), and \((1, 2)\) are critical points of the following function of three variables and classify them according to whether they are local minimums, local maximums or saddle points.

\[
f(x, y) = 5x^4 - 10x^3 + 17x^2 - 6yx^2 + 6yx - 12x + 5y^2 - 20y + 20.
\]
Answer:
The Hessian matrix is
\[
\begin{pmatrix}
  60x^2 + 34 - 12y - 60x & -12x + 6 \\
  -12x + 6 & -12x + 6
\end{pmatrix}
\]
Check its eigenvalues at the critical points. First consider the point \(\left(\frac{1}{7}, \frac{37}{20}\right)\). At this point, the Hessian matrix is
\[
\begin{pmatrix}
  -\frac{16}{7} & 0 \\
  0 & 10
\end{pmatrix}
\]
and its eigenvalues are \(-\frac{16}{7}, 10\). Therefore, there is a saddle point.

Next consider \((0, 2)\) at this point the Hessian matrix is
\[
\begin{pmatrix}
  10 & 6 \\
  6 & 10
\end{pmatrix}
\]
and the eigenvalues are 16, 4. Therefore, there is a local minimum at this point. There
is also a local minimum at the critical point, (1, 2).

6. Show the points \((\frac{1}{2}, -\frac{17}{8})\), \((0, -2)\), and \((1, -2)\) are critical points of the following function
of three variables and classify them according to whether they are local minimums,
local maximums or saddle points.
\[
f(x, y) = 4x^4 - 8x^3 - 4yx^2 + 4yx + 8x - 4x^2 + 4y^2 + 16y + 16. 
\]
Answer:
The Hessian matrix is \[
\begin{pmatrix}
48x^2 - 8 & -8y - 48x & -8x + 4 \\
-8x + 4 & -8 & 8 \\
-8x + 4 & 8 & 8 \\
\end{pmatrix}
\]. Check its eigenvalues at
the critical points. First consider the point \((\frac{1}{2}, -\frac{17}{8})\). This matrix is
\[
\begin{pmatrix}
-3 & 0 \\
0 & 8 \\
\end{pmatrix}
\] and its eigenvalues are −3, 8.
Next consider \((0, -2)\) at this point the Hessian matrix is
\[
\begin{pmatrix}
8 & 4 \\
4 & 8 \\
\end{pmatrix}
\] and the eigenvalues are 12, 4. Finally consider the point \((1, -2)\).
\[
\begin{pmatrix}
8 & -4 \\
-4 & 8 \\
\end{pmatrix}, 
\text{eigenvalues: } 12, 4.
\]
If the eigenvalues are both negative, then local max. If both positive, then local min.
Otherwise the test fails.

7. Find the critical points of the following function of three variables and classify them
according to whether they are local minimums, local maximums or saddle points.
\[
f(x, y, z) = \frac{1}{3} x^2 + \frac{22}{3} x + \frac{4}{3} - \frac{16}{3} yx - \frac{58}{3} y - \frac{4}{3} zx - \frac{46}{3} z + \frac{1}{3} y^2 - \frac{4}{3} zy - \frac{5}{3} z^2.
\]
Answer:
The critical point is at \((-2, 3, -5)\). The eigenvalues of the Hessian matrix at this point
are −6, −2, and 6.

8. Find the critical points of the following function of three variables and classify them
according to whether they are local minimums, local maximums or saddle points.
\[
f(x, y, z) = -\frac{5}{3} x^2 + \frac{2}{3} x - \frac{2}{3} + \frac{8}{3} yx + \frac{4}{3} y + \frac{14}{3} zx - \frac{28}{3} z - \frac{5}{3} y^2 + \frac{14}{3} zy - \frac{8}{3} z^2.
\]
Answer:
The eigenvalues are 4, −10, and −6 and the only critical point is \((1, 1, 0)\).

9. Find the critical points of the following function of three variables and classify them
according to whether they are local minimums, local maximums or saddle points.
\[
f(x, y, z) = -\frac{11}{3} x^2 + \frac{40}{3} x - \frac{56}{3} yx + \frac{10}{3} y - \frac{4}{3} zx + \frac{22}{3} z - \frac{11}{3} y^2 - \frac{4}{3} zy - \frac{5}{3} z^2.
\]

10. Find the critical points of the following function of three variables and classify them
according to whether they are local minimums, local maximums or saddle points.
\[
f(x, y, z) = -\frac{2}{3} x^2 + \frac{22}{3} x + \frac{37}{3} yx + \frac{11}{3} y - \frac{4}{3}zx - \frac{29}{3} z - \frac{5}{3} y^2 - \frac{3}{3} zy + \frac{7}{3} z^2.
\]

11. Show that if \(f\) has a critical point and some eigenvalue of the Hessian matrix is
positive, then there exists a direction in which when \(f\) is evaluated on the line through
the critical point having this direction, the resulting function of one variable has a
local minimum. State and prove a similar result in the case where some eigenvalue of
the Hessian matrix is negative.
12. Suppose \( \mu = 0 \) but there are negative eigenvalues of the Hessian at a critical point. Show by giving examples that the second derivative tests fails.

13. Show the points \((\frac{1}{2}, -\frac{2}{3})\), \((0, -5)\), and \((1, -5)\) are critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

\[
f(x, y) = 2x^4 - 4x^3 + 42x^2 + 8yx^2 - 8yx - 40x + 2y^2 + 20y + 50.
\]

14. Show the points \((1, -\frac{11}{2})\), \((0, -5)\), and \((2, -5)\) are critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

\[
f(x, y) = 4x^4 - 16x^3 - 4x^2 - 4yx^2 + 8yx + 40x + 4y^2 + 40y + 100.
\]

15. Show the points \((\frac{1}{2}, -\frac{2}{3})\), \((0, 0)\), and \((3, 0)\) are critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

\[
f(x, y) = 5x^4 - 30x^3 + 45x^2 + 6yx^2 - 18yx + 5y^2.
\]

16. Find the critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

\[
f(x, y, z) = \frac{10}{3}x^2 - \frac{44}{3}x + \frac{64}{3}y + \frac{10}{3}zx + \frac{16}{3}y + \frac{7}{3}z - \frac{20}{3}z + \frac{10}{3}y^2 + \frac{2}{3}yz + \frac{4}{3}z^2.
\]

17. Find the critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

\[
f(x, y, z) = -\frac{7}{3}x^2 - \frac{146}{3}x + \frac{83}{3} + \frac{16}{3}yx + \frac{4}{3}y - \frac{14}{3}zx + \frac{94}{3}z - \frac{7}{3}y^2 - \frac{14}{3}zy + \frac{8}{3}z^2.
\]

18. Find the critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

\[
f(x, y, z) = \frac{2}{3}x^2 + 4x + 75 - \frac{14}{3}yx - 38y - \frac{2}{3}zx - 2z + \frac{2}{3}y^2 - \frac{2}{3}zy - \frac{1}{3}z^2.
\]

19. Find the critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

\[
f(x, y, z) = 4zx - 30x + 510 - 2yx + 60y - 2zx - 70y + 4y^2 - 2zy + 4z^2.
\]

20. Show the critical points of the following function are points of the form, \((x, y, z) = (t, 2t^2 - 10t, -t^2 + 5t)\) for \(t \in \mathbb{R}\) and classify them as local minimums, local maximums or saddle points.

\[
f(x, y, z) = -\frac{1}{6}x^4 + \frac{5}{6}x^3 - \frac{25}{6}x^2 + \frac{10}{3}yx^2 - \frac{50}{3}yx + \frac{19}{3}zx^2 - \frac{95}{3}zx - \frac{5}{3}y^2 - \frac{10}{3}zy - \frac{1}{6}z^2.
\]

The verification that the critical points are of the indicated form is left for you.

The Hessian is

\[
\left(\begin{array}{cccc}
-2x^2 & \frac{10x - 25}{3} & \frac{20y + 45}{3} & \frac{20x - 50}{3} \\
\frac{20x - 50}{3} & \frac{10y - 10}{3} & \frac{45}{3} & \frac{10}{3} \\
\frac{20x - 50}{3} & \frac{45}{3} & \frac{10y - 10}{3} & \frac{10}{3} \\
\frac{10}{3} & \frac{10}{3} & \frac{10}{3} & -1
\end{array}\right)
\]

at a critical point it is

\[
\left(\begin{array}{cccc}
-\frac{1}{3}t^2 & \frac{40}{3}t - \frac{25}{3} & \frac{20}{3} (t) - \frac{50}{3} & \frac{45}{3} (t) - \frac{95}{3} \\
\frac{40}{3} (t) - \frac{50}{3} & \frac{10}{3} & \frac{45}{3} & \frac{10}{3} \\
\frac{20}{3} (t) - \frac{50}{3} & \frac{45}{3} & \frac{10}{3} & \frac{10}{3} \\
\frac{10}{3} & \frac{10}{3} & \frac{10}{3} & -1
\end{array}\right).
\]
The eigenvalues are
\[ 0, \frac{-2}{3} t^2 + \frac{10}{3} t - 6 + \frac{2}{3} \sqrt{(t^4 - 10t^3 + 493t^2 - 2340t + 2916)}, \]
and
\[ \frac{-2}{3} t^2 + \frac{10}{3} t - 6 - \frac{2}{3} \sqrt{(t^4 - 10t^3 + 493t^2 - 2340t + 2916)}. \]
If you graph these functions of \( t \) you find the second is always positive and the third is always negative. Therefore, all these critical points are saddle points.

21. Show the critical points of the following function are \((0, -3, 0), (2, -3, 0), \) and \((1, -3, -\frac{1}{3})\) and classify them as local minimums, local maximums or saddle points.

\[ f(x, y, z) = -\frac{3}{2} x^4 + 6x^3 - 6x^2 + zx^2 - 2zx - 2y^2 - 12y - 18 - \frac{3}{2} z^2. \]

The Hessian is
\[
\begin{pmatrix}
-12 + 36x + 2z - 18x^2 & 0 & -2 + 2x \\
0 & -4 & 0 \\
-2 + 2x & 0 & -3
\end{pmatrix}
\]

Now consider the critical point, \((1, -3, -\frac{1}{3})\). At this point the Hessian matrix equals
\[
\begin{pmatrix}
\frac{16}{3} & 0 & 0 \\
0 & -4 & 0 \\
0 & 0 & -3
\end{pmatrix},
\]

The eigenvalues are \(\frac{16}{3}, -3, -4\) and so this point is a saddle point.

Next consider the critical point, \((2, -3, 0)\). At this point the Hessian matrix is
\[
\begin{pmatrix}
-12 & 0 & 2 \\
0 & -4 & 0 \\
2 & 0 & -3
\end{pmatrix}
\]

The eigenvalues are \(-4, \frac{-15}{2} + \frac{1}{2} \sqrt{97}, \frac{-15}{2} - \frac{1}{2} \sqrt{97}\), all negative so at this point there is a local max.

Finally consider the critical point, \((0, -3, 0)\). At this point the Hessian is
\[
\begin{pmatrix}
-12 & 0 & -2 \\
0 & -4 & 0 \\
-2 & 0 & -3
\end{pmatrix}
\]
and the eigenvalues are the same as the above, all negative. Therefore, there is a local maximum at this point.

22. Show the critical points of the following function 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 minimums, local maximums or saddle points.

\[ f(x, y, z) = -2yzx^2 - 6yx - 4zx^2 - 12zx + y^2 + 2yz. \]

23. Show the critical points of the following function are \((0, -1, 0), (4, -1, 0), \) and \((2, -1, -12)\) and classify them as local minimums, local maximums or saddle points.

\[ f(x, y, z) = \frac{1}{2} x^4 - 4x^3 + 8x^2 - 3zx^2 + 12zx + 2y^2 + 4y + 2 + \frac{1}{2} z^2. \]
Lagrange multipliers are used to solve extremum problems for a function defined on a level set of another function. For example, suppose you want to maximize $xy$ given that $x + y = 4$. This is not too hard to do using methods developed earlier. Solve for one of the variables, say $y$, in the constraint equation, $x + y = 4$ to find $y = 4 - x$. Then the function to maximize is $f(x) = x(4 - x)$ and the answer is clearly $x = 2$. Thus the two numbers are $x = y = 2$.

This was easy because you could easily solve the constraint equation for one of the variables in terms of the other. Now what if you wanted to maximize $f(x, y, z) = xyz$ subject to the constraint that $x^2 + y^2 + z^2 = 4$? It is still possible to do this using similar techniques. Solve for one of the variables in the constraint equation, say $z$, substitute it into $f$, and then find where the partial derivatives equal zero to find candidates for the extremum. However, it seems you might encounter many cases and it does look a little fussy. However, sometimes you can’t solve the constraint equation for one variable in terms of the others. Also, what if you had many constraints? What if you wanted to maximize $f(x, y, z)$ subject to the constraints $x^2 + y^2 = 4$ and $z = 2x + 3y^2$. Things are clearly getting more involved and messy. It turns out that at an extremum, there is a simple relationship between the gradient of the function to be maximized and the gradient of the constraint function. This relation can be seen geometrically as in the following picture.

In the picture, the surface represents a piece of the level surface of $g(x, y, z) = 0$ and $f(x, y, z)$ is the function of three variables which is being maximized or minimized on the level surface and suppose the extremum of $f$ occurs at the point $(x_0, y_0, z_0)$. As shown above, $\nabla g(x_0, y_0, z_0)$ is perpendicular to the surface or more precisely to the tangent plane. However, if $x(t) = (x(t), y(t), z(t))$ is a point on a smooth curve which passes through $(x_0, y_0, z_0)$ when $t = t_0$, then the function, $h(t) = f(x(t), y(t), z(t))$ must have either a maximum or a minimum at the point, $t = t_0$. Therefore, $h'(t_0) = 0$. But this means

$$0 = h'(t_0) = \nabla f(x(t_0), y(t_0), z(t_0)) \cdot x'(t_0)$$

and since this holds for any such smooth curve, $\nabla f(x_0, y_0, z_0)$ is also perpendicular to the surface. This picture represents a situation in three dimensions and you can see that it is
intuitively clear that this implies \( \nabla f(x_0, y_0, z_0) \) is some scalar multiple of \( \nabla g(x_0, y_0, z_0) \). Thus
\[
\nabla f(x_0, y_0, z_0) = \lambda \nabla g(x_0, y_0, z_0)
\]
This \( \lambda \) is called a Lagrange multiplier after Lagrange who considered such problems in the 1700's.

Of course the above argument is at best only heuristic. It does not deal with the question of existence of smooth curves lying in the constraint surface passing through \((x_0, y_0, z_0)\). Nor does it consider all cases, being essentially confined to three dimensions. In addition to this, it fails to consider the situation in which there are many constraints. However, I think it is likely a geometric notion like that presented above which led Lagrange to formulate the method.

**Example 19.5.1** Maximize \( xyz \) subject to \( x^2 + y^2 + z^2 = 27 \).

Here \( f(x, y, z) = xyz \) while \( g(x, y, z) = x^2 + y^2 + z^2 - 27 \). Then \( \nabla g(x, y, z) = (2x, 2y, 2z) \) and \( \nabla f(x, y, z) = (yz, xz, xy) \). Then at the point which maximizes this function\(^1\),
\[
(yz, xz, xy) = \lambda (2x, 2y, 2z).
\]
Therefore, each of \( 2\lambda x^2, 2\lambda y^2, 2\lambda z^2 \) equals \( xyz \). It follows that at any point which maximizes \( xyz, |x| = |y| = |z| \). Therefore, the only candidates for the point where the maximum occurs are \((3,3,3),(-3,-3,3)(-3,3,3),\) etc. The maximum occurs at \((3,3,3)\) which can be verified by plugging in to the function which is being maximized.

**Example 19.5.2** Maximize \( f(x, y) = xy + y \) subject to the constraint, \( x^2 + y^2 \leq 1 \).

Here I know there is a maximum because the set is the closed circle, a closed and bounded set. Therefore, it is just a matter of finding it. Look for singular points on the interior of the circle. \( \nabla f(x, y) = (y, x + 1) = (0,0) \). There are no points on the interior of the circle where the gradient equals zero. Therefore, the maximum occurs on the boundary of the circle. That is the problem reduces to maximizing \( xy + y \) subject to \( x^2 + y^2 = 1 \). From the above,
\[
(y, x + 1) - \lambda (2x, 2y) = 0.
\]
Hence \( y^2 - 2\lambda xy = 0 \) and \( x(x + 1) - 2\lambda xy = 0 \) so \( y^2 = x(x + 1) \). Therefore from the constraint, \( x^2 + x (x + 1) = 1 \) and the solution is \( x = -1, x = \frac{1}{2} \). Then the candidates for a solution are \((-1,0), \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right), \left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right)\). Then
\[
f(-1,0) = 0, f \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right) = \frac{3\sqrt{3}}{4}, f \left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right) = -\frac{3\sqrt{3}}{4}.
\]
It follows the maximum value of this function is \( \frac{3\sqrt{3}}{4} \) and it occurs at \( \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right) \). The minimum value is \( -\frac{3\sqrt{3}}{4} \) and it occurs at \( \left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right) \).

This illustrates how to use the method of Lagrange multipliers to identify the extrema for a function defined on a closed and bounded set. You try and consider the boundary as a level curve or level surface and then use the method of Lagrange multipliers on it and look for singular points on the interior of the set.

There are no magic bullets here. It was still required to solve a system of nonlinear equations to get the answer. However, it does often help to do it this way.

\(^1\)There exists such a point because the sphere is closed and bounded.
The above generalizes to a general procedure which is described in the following major
Theorem. All correct proofs of this theorem will involve some appeal to the implicit function
theorem or to fundamental existence theorems from differential equations. A complete proof
is very fascinating but it will not come cheap. Good advanced calculus books will usually
give a correct proof. First here is a simple definition explaining one of the terms in the
statement of this theorem.

**Definition 19.5.3** Let \( A \) be an \( m \times n \) matrix. A submatrix is any matrix which can be
obtained from \( A \) by deleting some rows and some columns.

**Theorem 19.5.4** 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
\( g_i(x) = 0, \ i = 1, \ldots, m \) \hspace{1cm} (19.4)

and if some \( m \times m \) submatrix of
\[ Dg(x_0) \equiv \begin{pmatrix}
g_{1x_1}(x_0) & g_{1x_2}(x_0) & \cdots & g_{1x_n}(x_0) \\
\vdots & \vdots & & \vdots \\
g_{mx_1}(x_0) & g_{mx_2}(x_0) & \cdots & g_{mx_n}(x_0)
\end{pmatrix} \]
has nonzero determinant, then there exist scalars, \( \lambda_1, \ldots, \lambda_m \) such that
\[ \begin{pmatrix}
f_{x_1}(x_0) \\
\vdots \\
f_{x_n}(x_0)
\end{pmatrix} = \lambda_1 \begin{pmatrix}
g_{1x_1}(x_0) \\
\vdots \\
g_{1x_n}(x_0)
\end{pmatrix} + \cdots + \lambda_m \begin{pmatrix}
g_{mx_1}(x_0) \\
\vdots \\
g_{mx_n}(x_0)
\end{pmatrix} \hspace{1cm} (19.5) \]
holds.

To help remember how to use (19.5) it may be helpful to do the following. First write
the Lagrangian,
\[ L = f(x) - \sum_{i=1}^{m} \lambda_i g_i(x) \]
and then proceed to take derivatives with respect to each of the components of \( x \) and also
derivatives with respect to each \( \lambda_i \) and set all of these equations equal to 0. The formula
(19.5) is what results from taking the derivatives of \( L \) with respect to the components of
\( x \). When you take the derivatives with respect to the Lagrange multipliers, and set what
results equal to 0, you just pick up the constraint equations. This yields \( n + m \) equations
for the \( n + m \) unknowns, \( x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m \). Then you proceed to look for solutions to
these equations. Of course these might be impossible to find using methods of algebra, but
you just do your best and hope it will work out.

**Example 19.5.5** Minimize \( xyz \) subject to the constraints \( x^2 + y^2 + z^2 = 4 \) and \( x - 2y = 0 \).

Form the Lagrangian,
\[ L = xyz - \lambda (x^2 + y^2 + z^2 - 4) - \mu (x - 2y) \]
and proceed to take derivatives with respect to every possible variable, leading to the fol-
lowing system of equations.
\[
\begin{align*}
yz - 2\lambda x - \mu &= 0 \\
xz - 2\lambda y + 2\mu &= 0 \\
xy - 2\lambda z &= 0 \\
x^2 + y^2 + z^2 &= 4 \\
x - 2y &= 0
\end{align*}
\]
Now you have to find the solutions to this system of equations. In general, this could be very hard or even impossible. If $\lambda = 0$, then from the third equation, either $x$ or $y$ must equal 0. Therefore, from the first two equations, $\mu = 0$ also. If $\mu = 0$ and $\lambda \neq 0$, then from the first two equations, $xyz = 2\lambda x^2$ and $xyz = 2\lambda y^2$ and so either $x = y$ or $x = -y$, which requires that both $x$ and $y$ equal zero thanks to the last equation. But then from the fourth equation, $z = \pm 2$ and now this contradicts the third equation. Thus $\mu$ and $\lambda$ are either both equal to zero or neither one is and the expression, $xyz$ equals zero in this case. However, I know this is not the best value for a minimizer because I can take $x = 2\sqrt{\frac{2}{5}}, y = \sqrt{\frac{2}{5}}$, and $z = -1$. This satisfies the constraints and the product of these numbers equals a negative number. Therefore, both $\mu$ and $\lambda$ must be non zero. Now use the last equation eliminate $x$ and write the following system.

\[
\begin{align*}
5y^2 + z^2 &= 4 \\
y^2 - \lambda z &= 0 \\
yz - \lambda y + \mu &= 0 \\
yz - 4\lambda y - \mu &= 0
\end{align*}
\]

From the last equation, $\mu = (yz - 4\lambda y)$. Substitute this into the third and get

\[
\begin{align*}
5y^2 + z^2 &= 4 \\
y^2 - \lambda z &= 0 \\
yz - \lambda y + yz - 4\lambda y &= 0
\end{align*}
\]

$y = 0$ will not yield the minimum value from the above example. Therefore, divide the last equation by $y$ and solve for $\lambda$ to get $\lambda = (2/5)z$. Now put this in the second equation to conclude

\[
\begin{align*}
5y^2 + z^2 &= 4 \\
y^2 - (2/5)z^2 &= 0
\end{align*}
\]

a system which is easy to solve. Thus $y^2 = 8/15$ and $z^2 = 4/3$. Therefore, candidates for minima are \(\left(2\sqrt{\frac{8}{15}}, \sqrt{\frac{8}{15}}, \pm \sqrt{\frac{4}{3}}\right)\), and \(\left(-2\sqrt{\frac{8}{15}}, -\sqrt{\frac{8}{15}}, \pm \sqrt{\frac{4}{3}}\right)\), a choice of 4 points to check. Clearly the one which gives the smallest value is \(\left(2\sqrt{\frac{8}{15}}, \sqrt{\frac{8}{15}}, -\sqrt{\frac{4}{3}}\right)\) or \(\left(-2\sqrt{\frac{8}{15}}, -\sqrt{\frac{8}{15}}, -\sqrt{\frac{4}{3}}\right)\) and the minimum value of the function subject to the constraints is $-\frac{2}{5}\sqrt{30} - \frac{2}{3}\sqrt{3}$.

You should rework this problem first solving the second easy constraint for $x$ and then producing a simpler problem involving only the variables $y$ and $z$.

The method of Lagrange multipliers allows you to consider maximization of functions defined on closed and bounded sets. Recall that any continuous function defined on a closed and bounded set has a maximum and a minimum on the set. Candidates for the extremum on the interior of the set can be located by setting the gradient equal to zero. The consideration of the boundary can then sometimes be handled with the method of Lagrange multipliers.

**Example 19.5.6** Find the maximum and minimum values of the function, $f(x,y) = xy - x^2$ on the set, \(\{(x,y) : x^2 + 2xy + y^2 \leq 4\}\).

First, the only point where $\nabla f$ equals zero is $(x, y) = (0, 0)$ and this is in the desired set. In fact it is an interior point of this set. This takes care of the interior points. What about
those on the boundary $x^2 + 2xy + y^2 = 4$? The problem is to maximize $xy - x^2$ subject to the constraint, $x^2 + 2xy + y^2 = 4$. The Lagrangian is $xy - x^2 - \lambda (x^2 + 2xy + y^2 - 4)$ and this yields the following system.

$$
\begin{align*}
y - 2x - \lambda (2x + 2y) &= 0 \\
x - 2\lambda (x + y) &= 0 \\
2x^2 + 2xy + y^2 &= 4
\end{align*}
$$

From the first two equations,

$$
\begin{align*}
(2 + 2\lambda) x - (1 - 2\lambda) y &= 0 \\
(1 - 2\lambda) x - 2\lambda y &= 0
\end{align*}
$$

Since not both $x$ and $y$ equal zero, it follows

$$
\det \begin{pmatrix} 2 + 2\lambda & 2\lambda - 1 \\ 1 - 2\lambda & -2\lambda \end{pmatrix} = 0
$$

which yields

$$\lambda = 1/8$$

Therefore,

$$y = -\frac{3}{4} x$$

From the constraint equation,

$$2x^2 + 2x \left( -\frac{3}{4} x \right) + \left( -\frac{3}{4} x \right)^2 = 4$$

and so

$$x = \frac{8}{17}\sqrt{17} \text{ or } -\frac{8}{17}\sqrt{17}$$

Now from (19.6), the points of interest on the boundary of this set are

$$\left( \frac{8}{17}\sqrt{17}, -\frac{6}{17}\sqrt{17} \right), \text{ and } \left( -\frac{8}{17}\sqrt{17}, \frac{6}{17}\sqrt{17} \right). \tag{19.7}$$

$$f \left( \frac{8}{17}\sqrt{17}, -\frac{6}{17}\sqrt{17} \right) = \left( \frac{8}{17}\sqrt{17} \right) \left( -\frac{6}{17}\sqrt{17} \right) - \left( \frac{8}{17}\sqrt{17} \right)^2 = -\frac{112}{17}$$

$$f \left( -\frac{8}{17}\sqrt{17}, \frac{6}{17}\sqrt{17} \right) = \left( -\frac{8}{17}\sqrt{17} \right) \left( \frac{6}{17}\sqrt{17} \right) - \left( -\frac{8}{17}\sqrt{17} \right)^2 = -\frac{112}{17}$$

It follows the maximum value of this function on the given set occurs at $(0, 0)$ and is equal to zero and the minimum occurs at either of the two points in (19.7) and has the value $-112/17$. 

19.6 Exercises

1. Maximize \(2x + 3y - 6z\) subject to the constraint, \(x^2 + 2y^2 + 3z^2 = 9\).

2. Find the dimensions of the largest rectangle which can be inscribed in a circle of radius \(r\).

3. Maximize \(2x + y\) subject to the condition that \(\frac{x^2}{4} + \frac{y^2}{1} \leq 1\).

4. Maximize \(x + 2y\) subject to the condition that \(x^2 + \frac{y^2}{1} \leq 1\).

5. Maximize \(x + y\) subject to the condition that \(x^2 + \frac{y^2}{2} + z^2 \leq 1\).

6. Maximize \(x + y + z\) subject to the condition that \(x^2 + \frac{y^2}{4} + z^2 \leq 1\).

7. Find the points on \(y^2 = x\) which are closest to \((0, 0)\).

8. Find points on \(xy = 4\) farthest from \((0, 0)\) if any exist. If none exist, tell why. What does this say about the method of Lagrange multipliers?

9. A can is supposed to have a volume of \(36\pi\) cubic centimeters. Find the dimensions of the can which minimizes the surface area.

10. A can is supposed to have a volume of \(36\pi\) cubic centimeters. The top and bottom of the can are made of tin costing 4 cents per square centimeter and the sides of the can are made of aluminum costing 5 cents per square centimeter. Find the dimensions of the can which minimizes the cost.

11. 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.

12. Find the point, \((x, y, z)\) on the level surface, \(4x^2 + y^2 - z^2 = 1\) which is closest to \((0, 0, 0)\).

13. 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 \((0, 0, 0)\).

14. 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 \((0, 0, 0)\).

15. Find the point on the plane, \(2x + 3y + z = 4\) which is closest to the point \((1, 2, 3)\).

16. 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 \(\frac{\partial}{\partial x_j}(A_{ij}x_jx_i) = 2A_{ij}x_j\). Show that when you use the method of Lagrange multipliers to maximize the function, \(A_{ij}x_jx_i\) 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, \(A\).

17. Here are two lines. \(x = (1 + 2t, 2 + t, 3 + t)^T\) and \(x = (2 + s, 1 + 2s, 1 + 3s)^T\). Find points \(p_1\) on the first line and \(p_2\) on the second with the property that \(|p_1 - p_2|\) is at least as small as the distance between any other pair of points, one chosen on one line and the other on the other line.

18. Find the dimensions of the largest triangle which can be inscribed in a circle of radius \(r\).
19. Find the point on the intersection of $z = x^2 + y^2$ and $x + y + z = 1$ which is closest to $(0, 0, 0)$.

20. Minimize $4x^2 + y^2 + 9z^2$ subject to $x + y - z = 1$ and $x - 2y + z = 0$.

21. Minimize $xyz$ subject to the constraints $x^2 + y^2 + z^2 = r^2$ and $x - y = 0$.

22. Let $n$ be a positive integer. Find $n$ numbers whose sum is $8n$ and the sum of the squares is as small as possible.

23. Find the point on the level surface, $2x^2 + xy + z^2 = 16$ which is closest to $(0, 0, 0)$.

24. Find the point on $x^2 + 4y^2 + 9z^2 = 1$ closest to the plane $x + y + z = 10$.

25. Let $x_1, \cdots, x_5$ be 5 positive numbers. Maximize their product subject to the constraint that $x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 = 300$.

26. Let $f(x_1, \cdots, x_n) = x_1 x_2^{n-1} \cdots x_n^1$. Then $f$ achieves a maximum on the set, $S \equiv \{ x \in \mathbb{R}^n : \sum_{i=1}^n ix_i = 1 \text{ and each } x_i \geq 0 \}$.

If $x \in S$ is the point where this maximum is achieved, find $x_1/x_n$.

27. Let $(x, y)$ be a point on the ellipse, $x^2/a^2 + y^2/b^2 = 1$ which is in the first quadrant. Extend the tangent line through $(x, y)$ till it intersects the $x$ and $y$ axes and let $A(x, y)$ denote the area of the triangle formed by this line and the two coordinate axes. Find the maximum value of the area of this triangle as a function of $a$ and $b$.

28. Maximize $\prod_{i=1}^n x_i^2 \equiv x_1^2 x_2^{n-1} \cdots x_n^1$ 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 $x_i > 0$, 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.

29. Maximize $x^2 y^2$ subject to the constraint

$$\frac{x^{2p}}{p} + \frac{y^{2q}}{q} = r^2$$

where $p, q$ are real numbers larger than 1 which have the property that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Show the maximum is achieved when $x^{2p} = y^{2q}$ and equals $r^2$. Now conclude that if $x, y > 0$, then

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}$$

and there are values of $x$ and $y$ where this inequality is an equation.
The Riemann Integral On $\mathbb{R}^n$

20.0.1 Outcomes
1. Recall and define the Riemann integral.
2. Recall the relation between iterated integrals and the Riemann integral.
3. Evaluate double integrals over simple regions.
4. Evaluate multiple integrals over simple regions.
5. Use multiple integrals to calculate the volume and mass.

20.1 Methods For Double Integrals
This chapter is on the Riemann integral for a function of $n$ variables. It begins by introducing the basic concepts and applications of the integral. The proofs of the theorems involved are difficult and are left till the end. To begin with consider the problem of finding the volume under a surface of the form $z = f(x, y)$ where $f(x, y) \geq 0$ and $f(x, y) = 0$ for all $(x, y)$ outside of some bounded set. To solve this problem, consider the following picture.

In this picture, the volume of the little prism which lies above the rectangle $Q$ and the graph of the function would lie between $M_Q(f) v(Q)$ and $m_Q(f) v(Q)$ where

$$M_Q(f) \equiv \sup \{f(x) : x \in Q\}, \quad m_Q(f) \equiv \inf \{f(x) : x \in Q\},$$

(20.1)
and $v(Q)$ is defined as the area of $Q$. Now consider the following picture.

In this picture, it is assumed $f$ equals zero outside the circle and $f$ is a bounded nonnegative function. Then each of those little squares are the base of a prism of the sort in the previous picture and the sum of the volumes of those prisms should be the volume under the surface, $z = f(x, y)$. Therefore, the desired volume must lie between the two numbers,

$$\sum_Q M_Q(f) v(Q) \quad \text{and} \quad \sum_Q m_Q(f) v(Q)$$

where the notation, $\sum_Q M_Q(f) v(Q)$, means for each $Q$, take $M_Q(f)$, multiply it by the area of $Q$, $v(Q)$, and then add all these numbers together. Thus in $\sum_Q M_Q(f) v(Q)$, adds numbers which are at least as large as what is desired while in $\sum_Q m_Q(f) v(Q) \quad$ numbers are added which are at least as small as what is desired. Note this is a finite sum because by assumption, $f = 0$ except for finitely many $Q$, namely those which intersect the circle. The sum, $\sum_Q M_Q(f) v(Q)$ is called an upper sum, $\sum_Q m_Q(f) v(Q)$ is a lower sum, and the desired volume is caught between these upper and lower sums.

None of this depends in any way on the function being nonnegative. It also does not depend in any essential way on the function being defined on $\mathbb{R}^2$, although it is impossible to draw meaningful pictures in higher dimensional cases. To define the Riemann integral, it is necessary to first give a description of something called a grid. First you must understand that something like $[a, b] \times [c, d]$ is a rectangle in $\mathbb{R}^2$, having sides parallel to the axes. The situation is illustrated in the following picture.
(x, y) ∈ [a, b] × [c, d], means x ∈ [a, b] and also y ∈ [c, d] and the points which do this comprise the rectangle just as shown in the picture.

**Definition 20.1.1** For i = 1, 2, let \( \{\alpha_i^k\}_{k=-\infty}^{\infty} \) be points on \( \mathbb{R} \) which satisfy
\[
\lim_{k \to \infty} \alpha_i^k = \infty, \quad \lim_{k \to -\infty} \alpha_i^k = -\infty, \quad \alpha_i^k < \alpha_i^{k+1}.
\] (20.2)

For such sequences, define a grid on \( \mathbb{R}^2 \) denoted by \( \mathcal{G} \) or \( \mathcal{F} \) as the collection of rectangles of the form
\[
Q = [\alpha_1^k, \alpha_1^{k+1}] \times [\alpha_2^l, \alpha_2^{l+1}].
\] (20.3)

If \( \mathcal{G} \) is a grid, another grid, \( \mathcal{F} \) is a refinement of \( \mathcal{G} \) if every box of \( \mathcal{G} \) is the union of boxes of \( \mathcal{F} \).

For \( \mathcal{G} \) a grid, the expression,
\[
\sum_{Q \in \mathcal{G}} M_Q(f) \cdot v(Q)
\]
is called the upper sum associated with the grid, \( \mathcal{G} \) as described above in the discussion of the volume under a surface. Again, this means to take a rectangle from \( \mathcal{G} \) multiply \( M_Q(f) \) defined in (20.1) by its area, \( v(Q) \) and sum all these products for every \( Q \in \mathcal{G} \). The symbol,
\[
\sum_{Q \in \mathcal{G}} m_Q(f) \cdot v(Q),
\]
called a lower sum, is defined similarly. With this preparation it is time to give a definition of the Riemann integral of a function of two variables.

**Definition 20.1.2** Let \( f : \mathbb{R}^2 \to \mathbb{R} \) be a bounded function which equals zero for all \( (x, y) \) outside some bounded set. Then \( \int f \, dV \) is defined to be the unique number which lies between all upper sums and all lower sums. In the case of \( \mathbb{R}^2 \), it is common to replace the \( V \) with \( A \) and write this symbol as \( \int f \, dA \) where \( A \) stands for area.

This definition begs a difficult question. For which functions does there exist a unique number between all the upper and lower sums? This interesting and fundamental question is discussed in any advanced calculus book and may be seen in the appendix on the theory of the Riemann integral. It is a hard problem which was only solved in the first part of the twentieth century. When it was solved, it was also realized that the Riemann integral was not the right integral to use. First consider the question: How can the Riemann integral be computed? Consider the following picture in which \( f \) equals zero outside the rectangle \([a, b] \times [c, d] \).
It depicts a slice taken from the solid defined by \((x, y) : 0 \leq y \leq f(x, y)\). You see these when you look at a loaf of bread. If you wanted to find the volume of the loaf of bread, and you knew the volume of each slice of bread, you could find the volume of the whole loaf by adding the volumes of individual slices. It is the same here. If you could find the volume of the slice represented in this picture, you could add these up and get the volume of the solid. The slice in the picture corresponds to constant \(y\) and is assumed to be very thin, having thickness equal to \(h\). Denote the volume of the solid under the graph of \(z = f(x, y)\) on \([a, b] \times [c, y]\) by \(V(y)\). Then

\[
V(y + h) - V(y) \approx h \int_a^b f(x, y) \, dx
\]

where the integral is obtained by fixing \(y\) and integrating with respect to \(x\). It is hoped that the approximation would be increasingly good as \(h\) gets smaller. Thus, dividing by \(h\) and taking a limit, it is expected that

\[
V'(y) = \int_a^b f(x, y) \, dx, \quad V(c) = 0.
\]

Therefore, the volume of the solid under the graph of \(z = f(x, y)\) is given by

\[
\int_c^d \left( \int_a^b f(x, y) \, dx \right) \, dy
\]

(20.4)

but this was also the result of \(\int f \, dV\). Therefore, it is expected that this is a way to evaluate \(\int f \, dV\). Note what has been gained here. A hard problem, finding \(\int f \, dV\), is reduced to a sequence of easier problems. First do

\[
\int_a^b f(x, y) \, dx
\]

getting a function of \(y\), say \(F(y)\) and then do

\[
\int_c^d \left( \int_a^b f(x, y) \, dx \right) \, dy = \int_c^d F(y) \, dy.
\]

Of course there is nothing special about fixing \(y\) first. The same thing should be obtained from the integral,

\[
\int_a^b \left( \int_c^d f(x, y) \, dy \right) \, dx
\]

(20.5)

These expressions in (20.4) and (20.5) are called iterated integrals. They are tools for evaluating \(\int f \, dV\) which would be hard to find otherwise. In practice, the parenthesis is usually omitted in these expressions. Thus

\[
\int_a^b \left( \int_c^d f(x, y) \, dy \right) \, dx = \int_a^b \int_c^d f(x, y) \, dy \, dx
\]

and it is understood that you are to do the inside integral first and then when you have done it, obtaining a function of \(x\), you integrate this function of \(x\).

I have presented this for the case where \(f(x, y) \geq 0\) and the integral represents a volume, but there is no difference in the general case where \(f\) is not necessarily nonnegative.
Throughout, I have been assuming the notion of volume has some sort of independent meaning. This assumption is nonsense and is one of many reasons the above explanation does not rise to the level of a proof. It is only intended to make things plausible. A careful presentation which is not for the faint of heart is in an appendix.

Another aspect of this is the notion of integrating a function which is defined on some set, not on all $\mathbb{R}^2$. For example, suppose $f$ is defined on the set, $S \subseteq \mathbb{R}^2$. What is meant by $\int_S f \, dV$?

**Definition 20.1.3** Let $f : S \to \mathbb{R}$ where $S$ is a subset of $\mathbb{R}^2$. Then denote by $f_1$ the function defined by

$$f_1(x, y) \equiv \begin{cases} f(x, y) & \text{if } (x, y) \in S \\ 0 & \text{if } (x, y) \notin S \end{cases}.$$

Then

$$\int_S f \, dV \equiv \int f_1 \, dV.$$

**Example 20.1.4** Let $f(x, y) = x^2y + yx$ for $(x, y) \in [0, 1] \times [0, 2] \equiv R$. Find $\int_R f \, dV$.

This is done using iterated integrals like those defined above. Thus

$$\int_R f \, dV = \int_0^1 \int_0^2 (x^2y + yx) \, dy \, dx.$$

The inside integral yields

$$\int_0^2 (x^2y + yx) \, dy = 2x^2 + 2x$$

and now the process is completed by doing $\int_0^1$ to what was just obtained. Thus

$$\int_0^1 \int_0^2 (x^2y + yx) \, dy \, dx = \int_0^1 (2x^2 + 2x) \, dx = \frac{5}{3}.$$

If the integration is done in the opposite order, the same answer should be obtained.

$$\int_0^2 \int_0^1 (x^2y + yx) \, dx \, dy$$

$$\int_0^1 (x^2y + yx) \, dx = \frac{5}{6}y$$

Now

$$\int_0^2 \int_0^1 (x^2y + yx) \, dx \, dy = \int_0^2 \left(\frac{5}{6}y\right) \, dy = \frac{5}{3}.$$

If a different answer had been obtained it would have been a sign that a mistake had been made.

**Example 20.1.5** Let $f(x, y) = x^2y + yx$ for $(x, y) \in R$ where $R$ is the triangular region defined to be in the first quadrant, below the line $y = x$ and to the left of the line $x = 4$. Find $\int_R f \, dV$. 

Now from the above discussion,

\[ \int_{R} f \, dV = \int_{0}^{4} \int_{0}^{x} (x^2y + yx) \, dy \, dx \]

The reason for this is that \( x \) goes from 0 to 4 and for each fixed \( x \) between 0 and 4, \( y \) goes from 0 to the slanted line, \( y = x \). Thus \( y \) goes from 0 to \( x \). This explains the inside integral. Now \( \int_{0}^{x} (x^2y + yx) \, dy = \frac{1}{2}x^4 + \frac{1}{2}x^3 \) and so

\[ \int_{R} f \, dV = \int_{0}^{4} \left( \frac{1}{2}x^4 + \frac{1}{2}x^3 \right) \, dx = \frac{672}{5}. \]

What of integration in a different order? Let’s put the integral with respect to \( y \) on the outside and the integral with respect to \( x \) on the inside. Then

\[ \int_{R} f \, dV = \int_{0}^{4} \int_{y}^{4} (x^2y + yx) \, dx \, dy \]

For each \( y \) between 0 and 4, the variable \( x \), goes from \( y \) to 4.

\[ \int_{y}^{4} (x^2y + yx) \, dx = \frac{88}{3}y - \frac{1}{3}y^4 - \frac{1}{2}y^3 \]

Now

\[ \int_{R} f \, dV = \int_{0}^{4} \left( \frac{88}{3}y - \frac{1}{3}y^4 - \frac{1}{2}y^3 \right) \, dy = \frac{672}{5}. \]

Here is a similar example.

**Example 20.1.6** Let \( f(x, y) = x^2y \) for \((x, y) \in R\) where \( R \) is the triangular region defined to be in the first quadrant, below the line \( y = 2x \) and to the left of the line \( x = 4 \). Find \( \int_{R} f \, dV \).

Put the integral with respect to \( x \) on the outside first. Then

\[ \int_{R} f \, dV = \int_{0}^{4} \int_{0}^{x^2} (x^2y) \, dy \, dx \]
because for each $x \in [0, 4]$, $y$ goes from 0 to $2x$. Then

$$
\int_0^{2x} (x^2 y) \, dy = 2x^4
$$

and so

$$
\int_R f \, dV = \int_0^4 (2x^4) \, dx = \frac{2048}{5}
$$

Now do the integral in the other order. Here the integral with respect to $y$ will be on the outside. What are the limits of this integral? Look at the triangle and note that $x$ goes from 0 to 4 and so $2x = y$ goes from 0 to 8. Now for fixed $y$ between 0 and 8, where does $x$ go? It goes from the $x$ coordinate on the line $y = 2x$ which corresponds to this $y$ to 4. What is the $x$ coordinate on this line which goes with $y$? It is $x = y/2$. Therefore, the iterated integral is

$$
\int_0^8 \int_{y/2}^4 (x^2 y) \, dx \, dy.
$$

Now

$$
\int_{y/2}^4 (x^2 y) \, dx = \frac{64}{3} y - \frac{1}{24} y^4
$$

and so

$$
\int_R f \, dV = \int_0^8 \left( \frac{64}{3} y - \frac{1}{24} y^4 \right) \, dy = \frac{2048}{5}
$$

the same answer.

A few observations are in order here. In finding $\int_S f \, dV$ there is no problem in setting things up if $S$ is a rectangle. However, if $S$ is not a rectangle, the procedure always is agonizing. A good rule of thumb is that if what you do is easy it will be wrong. There are no shortcuts! There are no quick fixes which require no thought! Pain and suffering is inevitable and you must not expect it to be otherwise. Always draw a picture and then begin agonizing over the correct limits. Even when you are careful you will make lots of mistakes until you get used to the process.

Sometimes an integral can be evaluated in one order but not in another.

**Example 20.1.7** For $R$ as shown below, find $\int_R \sin \left( y^2 \right) \, dV$.

![Diagram of a region R with limits of integration](attachment:diagram.png)

Setting this up to have the integral with respect to $y$ on the inside yields

$$
\int_0^4 \int_{2x}^8 \sin \left( y^2 \right) \, dy \, dx.
$$

Unfortunately, there is no antiderivative in terms of elementary functions for $\sin \left( y^2 \right)$ so there is an immediate problem in evaluating the inside integral. It doesn’t work out so the
next step is to do the integration in another order and see if some progress can be made. This yields
\[ \int_0^8 \int_0^{y/2} \sin(y^2) \, dx \, dy = \int_0^8 \frac{y}{2} \sin(y^2) \, dy \]
and \[\int_0^8 \frac{y}{2} \sin(y^2) \, dy = -\frac{1}{4} \cos 64 + \frac{1}{4} \]
which you can verify by making the substitution, \( u = y^2 \). Thus
\[ \int_R \sin(y^2) \, dy = -\frac{1}{4} \cos 64 + \frac{1}{4} \]

This illustrates an important idea. The integral \( \int_R \sin(y^2) \, dV \) is defined as a number. It is the unique number between all the upper sums and all the lower sums. Finding it is another matter. In this case it was possible to find it using one order of integration but not the other. The iterated integral in this other order also is defined as a number but it can’t be found directly without interchanging the order of integration. Of course sometimes nothing you try will work out.

### 20.1.1 Density And Mass

Consider a two dimensional material. Of course there is no such thing but a flat plate might be modeled as one. The density \( \rho \) is a function of position and is defined as follows. Consider a small chunk of area, \( dV \) located at the point whose Cartesian coordinates are \((x, y)\). Then the mass of this small chunk of material is given by \( \rho(x, y) \, dV \). Thus if the material occupies a region in two dimensional space, \( U \), the total mass of this material would be
\[ \int_U \rho \, dV \]

In other words you integrate the density to get the mass. Now by letting \( \rho \) depend on position, you can include the case where the material is not homogeneous. Here is an example.

**Example 20.1.8** Let \( \rho(x, y) \) denote the density of the plane region determined by the curves \( \frac{1}{3}x + y = 2, x = 3y^2 \), and \( x = 9y \). Find the total mass if \( \rho(x, y) = y \).

You need to first draw a picture of the region, \( R \). A rough sketch follows.

![Sketch of the region](image)

This region is in two pieces, one having the graph of \( x = 9y \) on the bottom and the graph of \( x = 3y^2 \) on the top and another piece having the graph of \( x = 9y \) on the bottom and the graph of \( \frac{1}{3}x + y = 2 \) on the top. Therefore, in setting up the integrals, with the integral with respect to \( x \) on the outside, the double integral equals the following sum of iterated integrals.

\[ \int_0^3 \int_{\sqrt[3]{x/3}}^{\sqrt[3]{x/9}} y \, dy \, dx + \int_3^9 \int_{\sqrt{x/9}}^{\sqrt{x/3}} y \, dy \, dx \]
You notice it is not necessary to have a perfect picture, just one which is good enough to figure out what the limits should be. The dividing line between the two cases is \( x = 3 \) and this was shown in the picture. Now it is only a matter of evaluating the iterated integrals which in this case is routine and gives 1.

### 20.2 Exercises

1. Let \( \rho(x, y) \) denote the density of the plane region determined by the curves \( \frac{1}{4}x + y = 6, x = 4y^2, \) and \( x = 16y. \) Find the total mass if \( \rho(x, y) = y. \) Your answer should be \( \frac{1168}{75}. \)

2. Let \( \rho(x, y) \) denote the density of the plane region determined by the curves \( \frac{1}{5}x + y = 6, x = 5y^2, \) and \( x = 25y. \) Find the total mass if \( \rho(x, y) = y + 2x. \) Your answer should be \( \frac{1735}{3}. \)

3. Let \( \rho(x, y) \) denote the density of the plane region determined by the curves \( y = 3x, y = x, 3x + 3y = 9. \) Find the total mass if \( \rho(x, y) = y + 1. \) Your answer should be \( \frac{81}{12}. \)

4. Let \( \rho(x, y) \) denote the density of the plane region determined by the curves \( y = 3x, y = x, 4x + 2y = 8. \) Find the total mass if \( \rho(x, y) = y + 1. \)

5. Let \( \rho(x, y) \) denote the density of the plane region determined by the curves \( y = 3x, y = x, 2x + 2y = 4. \) Find the total mass if \( \rho(x, y) = x + 2y. \)

6. Let \( \rho(x, y) \) denote the density of the plane region determined by the curves \( y = 3x, y = x, 5x + 2y = 10. \) Find the total mass if \( \rho(x, y) = y + 1. \)

7. Find \( \int_0^4 \int_y^{\frac{1}{2}x} e^{2y} \, dx \, dy. \) Your answer should be \( e^4 - 1. \) You might need to interchange the order of integration.

8. Find \( \int_0^8 \int_y^{\frac{1}{2}x} e^{3y} \, dx \, dy. \)

9. Find \( \int_0^8 \int_y^{\frac{1}{2}x} \frac{1}{x} e^{3y} \, dx \, dy. \)

10. Find \( \int_0^1 \int_y^{\frac{1}{2}x} \frac{1}{x} e^{3y} \, dx \, dy. \)

11. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. \( \int_0^4 \int_0^{3x} xy^3 \, dx \, dy. \) Your answer for the iterated integral should be \( \int_0^{12} \int_0^{\frac{3}{4}x} xy^3 \, dy \, dx. \)

12. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. \( \int_0^3 \int_0^{3y} xy^3 \, dx \, dy. \)

13. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. \( \int_0^2 \int_0^{3y} xy^2 \, dx \, dy. \)

14. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. \( \int_0^3 \int_0^{3y} xy^2 \, dx \, dy. \)

15. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. \( \int_0^1 \int_0^{3y} xy^2 \, dx \, dy. \)

16. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. \( \int_0^5 \int_0^{3y} xy^2 \, dx \, dy. \)
17. Find $\int_0^{\frac{1}{4} \pi} \int_{\frac{1}{4} \pi}^y \sin \frac{y}{y} \, dy \, dx$. Your answer should be $\frac{1}{2}$.

18. Find $\int_0^{\frac{1}{4} \pi} \int_{\frac{1}{4} \pi}^y \sin \frac{y}{y} \, dy \, dx$.

19. Find $\int_0^{\pi} \int_{\frac{1}{4} \pi}^y \sin \frac{y}{y} \, dy \, dx$.

20. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_{-3}^3 \int_{-y}^y x^2 \, dy \, dx$.
Your answer for the iterated integral should be $\int_0^3 \int_{-y}^y x^2 \, dy \, dx + \int_{-3}^0 \int_{-y}^y x^2 \, dy \, dx$. This is a very interesting example which shows that iterated integrals have a life of their own, not just as a method for evaluating double integrals.

21. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_{-2}^2 \int_{-x}^x x^2 \, dy \, dx$.

### 20.3 Methods For Triple Integrals

#### 20.3.1 Definition Of The Integral

The integral of a function of three variables is defined similar to the integral of a function of two variables.

**Definition 20.3.1** For $i = 1, 2, 3$ let $\{a_k^i\}_{k=-\infty}^{\infty}$ be points on $\mathbb{R}$ which satisfy

$\lim_{k \to \infty} a_k^i = \infty, \quad \lim_{k \to -\infty} a_k^i = -\infty, \quad a_k^i < a_{k+1}^i$. \hfill (20.6)

For such sequences, define a grid on $\mathbb{R}^3$ denoted by $\mathcal{G}$ or $\mathcal{F}$ as the collection of boxes of the form

$Q = [a_k^1, a_{k+1}^1] \times [a_k^2, a_{k+1}^2] \times [a_k^3, a_{k+1}^3]$. \hfill (20.7)

If $\mathcal{G}$ is a grid, $\mathcal{F}$ is called a refinement of $\mathcal{G}$ if every box of $\mathcal{G}$ is the union of boxes of $\mathcal{F}$.

For $\mathcal{G}$ a grid,

$$\sum_{Q \in \mathcal{G}} M_Q(f) \, v(Q)$$

is the upper sum associated with the grid, $\mathcal{G}$ where

$$M_Q(f) \equiv \sup \{f(\mathbf{x}) : \mathbf{x} \in Q\}$$

and if $Q = [a, b] \times [c, d] \times [e, f]$, then $v(Q)$ is the volume of $Q$ given by $(b - a)(d - c)(f - e)$.

Letting

$$m_Q(f) \equiv \inf \{f(\mathbf{x}) : \mathbf{x} \in Q\}$$

the lower sum associated with this partition is

$$\sum_{Q \in \mathcal{G}} m_Q(f) \, v(Q),$$

With this preparation it is time to give a definition of the Riemann integral of a function of three variables. This definition is just like the one for a function of two variables.
Definition 20.3.2 Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ be a bounded function which equals zero outside of some bounded subset of $\mathbb{R}^3$. $\int f \, dV$ is defined as the unique number between all the upper sums and lower sums.

As in the case of a function of two variables there are all sorts of mathematical questions which are dealt with later.

The way to think of integrals is as follows. Located at a point $x$, there is an “infinitesimal” chunk of volume, $dV$, multiplying it by $f(x)$ and then adding up all such products. Upper sums are too large and lower sums are too small but the unique number between all the lower and upper sums is just right and corresponds to the notion of adding up all the $f(x) \, dV$. Even the notation is suggestive of this concept of sum. It is a long thin $S$ denoting sum. This is the fundamental concept for the integral in any number of dimensions and all the definitions and technicalities are designed to give precision and mathematical respectability to this notion.

To consider how to evaluate triple integrals, imagine a sum of the form $\sum_{ijk} a_{ijk}$ where there are only finitely many choices for $i, j, k$ and the symbol means you simply add up all the $a_{ijk}$. By the commutative law of addition, these may be added systematically in the form, $\sum_k \sum_j \sum_i a_{ijk}$. A similar process is used to evaluate triple integrals and since integrals are like sums, you might expect it to be valid. Specifically,

$$\int f \, dV = \int \int \int f(x, y, z) \, dx \, dy \, dz.$$ 

In words, sum with respect to $x$ and then sum what you get with respect to $y$ and finally, with respect to $z$. Of course this should hold in any other order such as

$$\int f \, dV = \int \int \int f(x, y, z) \, dz \, dy \, dx.$$ 

This is proved in an appendix\(^1\).

Having discussed double and triple integrals, the definition of the integral of a function of $n$ variables is accomplished in the same way.

Definition 20.3.3 For $i = 1, \ldots, n$, let $\{\alpha_{k}^{i}\}_{k=\infty}^{-\infty}$ be points on $\mathbb{R}$ which satisfy

$$\lim_{k \rightarrow \infty} \alpha_{k}^{i} = \infty, \quad \lim_{k \rightarrow -\infty} \alpha_{k}^{i} = -\infty, \quad \alpha_{k}^{i} < \alpha_{k+1}^{i}. \quad (20.8)$$

For such sequences, define a grid on $\mathbb{R}^n$ denoted by $\mathcal{G}$ or $\mathcal{F}$ as the collection of boxes of the form

$$Q = \prod_{i=1}^{n} [\alpha_{j_{i}}^{i}, \alpha_{j_{i}+1}^{i}]. \quad (20.9)$$

If $\mathcal{G}$ is a grid, $\mathcal{F}$ is called a refinement of $\mathcal{G}$ if every box of $\mathcal{G}$ is the union of boxes of $\mathcal{F}$.

Definition 20.3.4 Let $f$ be a bounded function which equals zero off a bounded set, $D$, and let $\mathcal{G}$ be a grid. For $Q \in \mathcal{G}$, define

$$M_Q(f) \equiv \sup \{ f(x) : x \in Q \}, \quad m_Q(f) \equiv \inf \{ f(x) : x \in Q \}. \quad (20.10)$$

Also define for $Q$ a box, the volume of $Q$, denoted by $v(Q)$ by

$$v(Q) \equiv \prod_{i=1}^{n} (b_i - a_i), \quad Q \equiv \prod_{i=1}^{n} [a_i, b_i].$$

\(^1\)All of these fundamental questions about integrals can be considered more easily in the context of the Lebesgue integral. However, this integral is more abstract than the Riemann integral.
Now define upper sums, $U_{G}(f)$ and lower sums, $L_{G}(f)$ with respect to the indicated grid, by the formulas

$$U_{G}(f) \equiv \sum_{Q \in G} M_{Q}(f) v(Q), \quad L_{G}(f) \equiv \sum_{Q \in G} m_{Q}(f) v(Q).$$

Then a function of $n$ variables is Riemann integrable if there is a unique number between all the upper and lower sums. This number is the value of the integral.

In this book most integrals will involve no more than three variables. However, this does not mean an integral of a function of more than three variables is unimportant. Therefore, I will begin to refer to the general case when theorems are stated.

**Definition 20.3.5** For $E \subseteq \mathbb{R}^{n}$,

$$\mathcal{X}_{E}(x) \equiv \begin{cases} 1 & \text{if } x \in E, \\ 0 & \text{if } x \notin E. \end{cases}$$

Define $\int_{E} f \, dV \equiv \int \mathcal{X}_{E} f \, dV$ when $f \mathcal{X}_{E} \in \mathcal{R}(\mathbb{R}^{n})$.

**20.3.2 Iterated Integrals**

As before, the integral is often computed by using an iterated integral.

**Example 20.3.6** Find $\int_{2}^{3} \int_{3}^{x} (x - y) \, dy \, dx$.

The inside integral yields $\int_{3}^{x} (x - y) \, dy = x^{2} - 4xy + 3y^{2}$. Next this must be integrated with respect to $y$ to give $\int_{3}^{x} (x^{2} - 4xy + 3y^{2}) \, dy = -3x^{2} + 18x - 27$. Finally the third integral gives

$$\int_{2}^{3} \left( -3x^{2} + 18x - 27 \right) \, dx = -1.$$

**Example 20.3.7** Find $\int_{0}^{\pi} \int_{0}^{\sqrt{y^{2} + z}} \cos(x + y) \, dz \, dx$.

The inside integral is $\int_{0}^{\sqrt{y^{2} + z}} \cos(x + y) \, dx = 2 \cos z \sin y \cos y + 2 \sin z \cos^{2} y - \sin z - \sin y$. Now this has to be integrated.

$$\int_{0}^{\sqrt{y^{2} + z}} \cos(x + y) \, dx \, dz = \int_{0}^{\sqrt{y^{2} + z}} \left( 2 \cos z \sin y \cos y + 2 \sin z \cos^{2} y - \sin z - \sin y \right) \, dz$$

$$= -1 - 16 \cos^{2} y + 20 \cos^{3} y - 5 \cos y - 3 (\sin y) y + 2 \cos^{2} y.$$ 

Finally, this last expression must be integrated from 0 to $\pi$. Thus

$$\int_{0}^{\pi} \int_{0}^{\sqrt{y^{2} + z}} \cos(x + y) \, dx \, dz \, dy$$

$$= \int_{0}^{\pi} \left( -1 - 16 \cos^{2} y + 20 \cos^{3} y - 5 \cos y - 3 (\sin y) y + 2 \cos^{2} y \right) \, dy$$

$$= -3 \pi$$

**Example 20.3.8** Here is an iterated integral: $\int_{0}^{2} \int_{0}^{\sqrt{3 - x^{2}}} \int_{0}^{\pi} \sin(x + y) \, dz \, dy \, dx$. Write as an iterated integral in the order $dz \, dx \, dy$. 


The inside integral is just a function of $x$ and $y$. (In fact, only a function of $x$.) The order of the last two integrals must be interchanged. Thus the iterated integral which needs to be done in a different order is

$$\int_0^2 \int_0^{x^3 - \frac{3}{2}x} f(x, y) \, dy \, dx.$$ 

As usual, it is important to draw a picture and then go from there.

Thus this double integral equals

$$\int_0^3 \int_0^{\frac{3}{2}(3-y)} f(x, y) \, dx \, dy.$$ 

Now substituting in for $f(x, y)$,

$$\int_0^3 \int_0^{\frac{3}{2}(3-y)} \int_0^x f(x, y) \, dz \, dx \, dy.$$ 

**Example 20.3.9** Find the volume of the bounded region determined by $3y + 3z = 2, x = 16 - y^2, y = 0, x = 0$.

In the $yz$ plane, the following picture corresponds to $x = 0$.

Therefore, the outside integrals taken with respect to $z$ and $y$ are of the form $\int_0^\frac{3}{2} \int_0^{\frac{3}{2} - y} dz \, dy$ and now for any choice of $(y, z)$ in the above triangular region, $x$ goes from 0 to $16 - y^2$.

Therefore, the iterated integral is

$$\int_0^\frac{3}{2} \int_0^{\frac{3}{2} - y} 16 - y^2 \, dx \, dz \, dy = \frac{860}{243}.$$ 

**Example 20.3.10** Find the volume of the region determined by the intersection of the two cylinders, $x^2 + y^2 \leq 9$ and $y^2 + z^2 \leq 9$.

The first listed cylinder intersects the $xy$ plane in the disk, $x^2 + y^2 \leq 9$. What is the volume of the three dimensional region which is between this disk and the two surfaces, $z = \sqrt{9 - y^2}$ and $z = -\sqrt{9 - y^2}$? An iterated integral for the volume is

$$\int_{-3}^3 \int_{-\sqrt{9-y^2}}^{\sqrt{9-y^2}} \int_{-\sqrt{9-y^2}}^{\sqrt{9-y^2}} dz \, dx \, dy = 144.$$
One of the cylinders is parallel to the $z$ axis, $x^2 + y^2 \leq 9$ and the other is parallel to the $x$ axis, $y^2 + z^2 \leq 9$. I did not need to be able to draw such a nice picture in order to work this problem. This is the key to doing these. Draw pictures in two dimensions and reason from the two dimensional pictures rather than attempt to wax artistic and consider all three dimensions at once. These problems are hard enough without making them even harder by attempting to be an artist.

### 20.3.3 Mass And Density

As an example of the use of triple integrals, consider a solid occupying a set of points, $U \subseteq \mathbb{R}^3$ having density $\rho$. Thus $\rho$ is a function of position and the total mass of the solid equals

$$\int_U \rho \, dV.$$ 

This is just like the two dimensional case. The mass of an infinitesimal chunk of the solid located at $x$ would be $\rho(x) \, dV$ and so the total mass is just the sum of all these, $\int_U \rho(x) \, dV$.

**Example 20.3.11** Find the volume of $R$ where $R$ is the bounded region formed by the plane $\frac{1}{5}x + y + \frac{1}{5}z = 1$ and the planes $x = 0, y = 0, z = 0$.

When $z = 0$, the plane becomes $\frac{1}{5}x + y = 1$. Thus the intersection of this plane with the $xy$ plane is this line shown in the following picture.

```
```

Therefore, the bounded region is between the triangle formed in the above picture by the $x$ axis, the $y$ axis and the above line and the surface given by $\frac{1}{5}x + y + \frac{1}{5}z = 1$ or $z = 5 \left(1 - \left(\frac{1}{5}x + y\right)\right) = 5 - x - 5y$. Therefore, an iterated integral which yields the volume is

$$\int_0^5 \int_0^{1 - \frac{1}{5}x} \int_0^{5-x-5y} dz \, dy \, dx = \frac{25}{6}.$$
Example 20.3.12 Find the mass of the bounded region, $R$ formed by the plane $\frac{1}{3}x + \frac{1}{3}y + \frac{1}{5}z = 1$ and the planes $x = 0, y = 0, z = 0$ if the density is $\rho(x, y, z) = z$.

This is done just like the previous example except in this case there is a function to integrate. Thus the answer is

$$\int_0^3 \int_0^{3-x} \int_0^{5-\frac{2}{3}x - \frac{4}{3}y} z \, dz \, dy \, dx = \frac{75}{8}.$$ 

Example 20.3.13 Find the total mass of the bounded solid determined by $z = 9 - x^2 - y^2$ and $x, y, z \geq 0$ if the mass is given by $\rho(x, y, z) = z$.

When $z = 0$ the surface, $z = 9 - x^2 - y^2$ intersects the xy plane in a circle of radius 3 centered at $(0, 0)$. Since $x, y \geq 0$, it is only a quarter of a circle of interest, the part where both these variables are nonnegative. For each $(x, y)$ inside this quarter circle, $z$ goes from 0 to $9 - x^2 - y^2$. Therefore, the iterated integral is of the form,

$$\int_0^3 \int_0^{\sqrt{9-x^2}} \int_0^{9-x^2-y^2} z \, dz \, dy \, dx = \frac{243}{8}\pi.$$ 

Example 20.3.14 Find the volume of the bounded region determined by $x \geq 0, y \geq 0, z \geq 0,$ and $\frac{1}{7}x + y + \frac{1}{4}z = 1$, and $x + \frac{1}{7}y + \frac{1}{4}z = 1$.

When $z = 0$, the plane $\frac{1}{7}x + y + \frac{1}{4}z = 1$ intersects the xy plane in the line whose equation is

$$\frac{1}{7}x + y = 1$$

while the plane, $x + \frac{1}{7}y + \frac{1}{4}z = 1$ intersects the xy plane in the line whose equation is

$$x + \frac{1}{7}y = 1.$$ 

Furthermore, the two planes intersect when $x = y$ as can be seen from the equations, $x + \frac{1}{7}y = 1 - \frac{1}{7}$ and $\frac{1}{7}x + y = 1 - \frac{1}{7}$ which imply $x = y$. Thus the two dimensional picture to look at is depicted in the following picture.

You see in this picture, the base of the region in the xy plane is the union of the two triangles, $R_1$ and $R_2$. For $(x, y) \in R_1$, $z$ goes from 0 to what it needs to be to be on the
plane, $\frac{1}{7}x + y + \frac{1}{4}z = 1$. Thus $z$ goes from 0 to $4 \left(1 - \frac{1}{7}x - y\right)$. Similarly, on $R_2$, $z$ goes from 0 to $4 \left(1 - \frac{1}{7}y - x\right)$. Therefore, the integral needed is

$$
\int_{R_1} \int_0^4 \left(1 - \frac{1}{7}x - y\right) dz \, dV + \int_{R_2} \int_0^4 \left(1 - \frac{1}{7}y - x\right) dz \, dV
$$

and now it only remains to consider $\int_{R_1} dV$ and $\int_{R_2} dV$. The point of intersection of these lines shown in the above picture is $(\frac{7}{8}, \frac{3}{8})$ and so an iterated integral is

$$
\int_0^{7/8} \int_{1-x/7}^4 \int_0^{4(1-\frac{1}{7}x-y)} dz \, dy \, dx + \int_0^{7/8} \int_{y}^1 \int_0^{4(1-\frac{1}{7}x-y)} dz \, dx \, dy = \frac{7}{6}.
$$

## 20.4 Exercises With Answers

The evaluation of integrals by setting up appropriate iterated integrals and then evaluating these requires a lot of practice. Therefore, I have included exercises with answers. Each of these exercises corresponds to one which does not have answers in the next section.

1. Evaluate the integral $\int_4^7 \int_5^3 x \, dy \, dz \, dx$
   
   Answer: $-\frac{3417}{2}$

2. Find $\int_0^4 \int_0^{2-5x} \int_0^{4-2x-y} (2x) \, dz \, dy \, dx$
   
   Answer: $-\frac{2464}{3}$

3. Find $\int_0^2 \int_0^{2-5x} \int_0^{1-4x-3y} (2x) \, dz \, dy \, dx$
   
   Answer: $-\frac{196}{3}$

4. Evaluate the integral $\int_5^8 \int_4^3 \int_5^x (x - y) \, dz \, dy \, dx$
   
   Answer: $\frac{114607}{8}$

5. Evaluate the integral $\int_0^\pi \int_0^y \int_0^{y+z} \cos(x + y) \, dx \, dz \, dy$
   
   Answer: $-4\pi$

6. Evaluate the integral $\int_0^\pi \int_0^y \int_0^{y+z} \sin(x + y) \, dx \, dz \, dy$
   
   Answer: $-\frac{19}{4}$

7. Fill in the missing limits. $\int_0^1 \int_0^z \int_0^y f(x, y, z) \, dx \, dy \, dz = \int_0^y \int_0^z f(x, y, z) \, dx \, dz \, dy$,
   
   $\int_0^1 \int_0^z \int_0^y f(x, y, z) \, dx \, dy \, dz = \int_0^z \int_0^y f(x, y, z) \, dy \, dx \, dz$,

   $\int_0^1 \int_0^z \int_0^y f(x, y, z) \, dx \, dy \, dz = \int_0^z \int_0^y f(x, y, z) \, dy \, dx \, dz$,

   $\int_0^1 \int_0^y \int_0^{y+z} f(x, y, z) \, dx \, dy \, dz = \int_0^y \int_0^{y+z} f(x, y, z) \, dx \, dz \, dy$,
Find the volume of
\[ \int_0^2 \int_0^{x/2} \int_0^{x/4} f(x, y, z) \, dz \, dy \, dx. \]

Answer:
\[ \int_0^2 \int_0^{x/2} \int_0^{x/4} f(x, y, z) \, dz \, dy \, dx = \int_0^2 \int_0^1 \left( \int_0^{x/4} f(x, y, z) \, dz \right) \, dy \, dx. \]

\[ \int_0^2 \int_0^1 \left( \int_0^{x/4} f(x, y, z) \, dz \right) \, dy \, dx = \int_0^2 \int_0^1 \frac{f(x, y, z)}{4} \, dy \, dx. \]

\[ \int_0^2 \int_0^1 \frac{f(x, y, z)}{4} \, dy \, dx = \int_0^2 \int_0^1 \frac{f(x, y, z)}{4} \, dy \, dx. \]

8. Find the volume of the bounded region determined by \( y = \frac{1}{5}x + \frac{1}{4} \) and the planes \( x = 0, y = 0, z = 0 \).

Answer:
\[ \int_0^2 \int_0^1 \int_0^{y/2} f(x, y, z) \, dz \, dy \, dx = \int_0^2 \int_0^1 \left( \int_0^{y/2} f(x, y, z) \, dz \right) \, dy \, dx. \]

9. Find the volume of the bounded region formed by the planes \( y = \frac{1}{5}x + \frac{1}{4}z = 1 \) and the planes \( x = 0, y = 0, z = 0 \).

Answer:
\[ \int_0^2 \int_0^1 \int_0^{y/2} f(x, y, z) \, dz \, dy \, dx = \int_0^2 \int_0^1 \left( \int_0^{y/2} f(x, y, z) \, dz \right) \, dy \, dx. \]

10. Find the mass of the bounded region, \( R \) formed by the plane \( \frac{1}{4}x + \frac{1}{2}y + \frac{1}{4}z = 1 \) and the planes \( x = 0, y = 0, z = 0 \) if the density is \( \rho(x, y, z) = y \)

Answer:
\[ \int_0^2 \int_0^1 \int_0^{y/2} f(x, y, z) \, dz \, dy \, dx = \int_0^2 \int_0^1 \left( \int_0^{y/2} f(x, y, z) \, dz \right) \, dy \, dx. \]

11. Find the mass of the bounded region, \( R \) formed by the plane \( \frac{1}{2}x + \frac{1}{2}y + \frac{1}{2}z = 1 \) and the planes \( x = 0, y = 0, z = 0 \) if the density is \( \rho(x, y, z) = z^2 \)

Answer:
\[ \int_0^2 \int_0^1 \int_0^{y/2} f(x, y, z) \, dz \, dy \, dx = \int_0^2 \int_0^1 \left( \int_0^{y/2} f(x, y, z) \, dz \right) \, dy \, dx. \]

12. Here is an iterated integral: \( \int_0^3 \int_0^{x^2} dz \, dy \). Write as an iterated integral in the following orders: \( dz \, dy \), \( dx \, dz \, dy \), \( dx \, dy \, dz \), \( dy \, dz \, dx \).

Answer:
\[ \int_0^3 \int_0^{x^2} dz \, dy = \int_0^3 \int_0^{x^2} dz \, dy = \int_0^3 \int_0^{x^2} dz \, dy = \int_0^3 \int_0^{x^2} dz \, dy. \]

13. Find the volume of the bounded region determined by \( 5y + 2z = 4, x = 4 - y^2, y = 0, x = 0 \).

Answer:
\[ \int_0^2 \int_0^1 \int_0^{y^2} f(x, y, z) \, dz \, dx \, dy = \int_0^2 \int_0^1 \left( \int_0^{y^2} f(x, y, z) \, dz \right) \, dx \, dy. \]

\[ \int_0^2 \int_0^1 \left( \int_0^{y^2} f(x, y, z) \, dz \right) \, dx \, dy = \int_0^2 \int_0^1 \frac{1168}{317} \, dx \, dy. \]
14. Find the volume of the bounded region determined by $4y + 3z = 3, x = 4 - y^2, y = 0, x = 0$.
   Answer:
   $$
   \int_0^3 \int_0^{1-\frac{4}{y}} \int_0^{4-y^2} \, dz \, dx \, dy = \frac{375}{256}
   $$

15. Find the volume of the bounded region determined by $3y + z = 3, x = 4 - y^2, y = 0, x = 0$.
   Answer:
   $$
   \int_0^1 \int_0^{3-3y} \int_0^{1-y^2} \, dx \, dz \, dy = \frac{23}{4}
   $$

16. Find the volume of the region bounded by $x^2 + y^2 = 16, z = 3x, z = 0$, and $x \geq 0$.
   Answer:
   $$
   \int_0^4 \int_0^{\sqrt{16-x^2}} \int_0^{3x} \, dy \, dz \, dx = 128
   $$

17. Find the volume of the region bounded by $x^2 + y^2 = 25, z = 2x, z = 0$, and $x \geq 0$.
   Answer:
   $$
   \int_0^5 \int_0^{\sqrt{25-x^2}} \int_0^{2x} \, dy \, dz \, dx = \frac{500}{3}
   $$

18. Find the volume of the region determined by the intersection of the two cylinders, $x^2 + y^2 \leq 9$ and $y^2 + z^2 \leq 9$.
   Answer:
   $$
   8 \int_0^3 \int_0^{\sqrt{9-y^2}} \int_0^{\sqrt{9-y^2}} \, dz \, dx \, dy = 144
   $$

19. Find the total mass of the bounded solid determined by $z = a^2 - x^2 - y^2$ and $x, y, z \geq 0$
   if the mass is given by $\rho(x, y, z) = z$.
   Answer:
   $$
   \int_0^4 \int_0^{\sqrt{16-x^2}} \int_0^{16-x^2-y^2} \, \rho(x, y, z) \, dz \, dy \, dx = \frac{512}{3}\pi
   $$

20. Find the total mass of the bounded solid determined by $z = a^2 - x^2 - y^2$ and $x, y, z \geq 0$
   if the mass is given by $\rho(x, y, z) = x + 1$.
   Answer:
   $$
   \int_0^5 \int_0^{\sqrt{25-x^2}} \int_0^{25-x^2-y^2} \, \rho(x, y, z) \, dz \, dy \, dx = \frac{625}{8}\pi + \frac{1250}{3}
   $$

21. Find the volume of the region bounded by $x^2 + y^2 = 9, z = 0, z = 5 - y$
   Answer:
   $$
   \int_{-3}^3 \int_{-\sqrt{9-x^2}}^{\sqrt{9-x^2}} \int_0^{5-y} \, dz \, dy \, dx = 45\pi
   $$

22. Find the volume of the bounded region determined by $x \geq 0, y \geq 0, z \geq 0$, and
   $\frac{1}{2}x + y + \frac{1}{2}z = 1$, and $x + \frac{1}{2}y + \frac{1}{2}z = 1$.
   Answer:
   $$
   \int_0^\frac{2}{3} \int_0^{1-\frac{1}{2}x} \int_0^{2-x-2y} \, dz \, dy \, dx + \int_0^\frac{2}{3} \int_0^{1-\frac{1}{2}y} \int_0^{2-2x-y} \, dz \, dx \, dy = \frac{4}{9}
   $$

23. Find the volume of the bounded region determined by $x \geq 0, y \geq 0, z \geq 0$, and
   $\frac{1}{2}x + y + \frac{1}{2}z = 1$, and $x + \frac{1}{2}y + \frac{1}{2}z = 1$.
   Answer:
   $$
   \int_0^\frac{2}{3} \int_0^{1-\frac{1}{2}x} \int_0^{3-\frac{3}{2}x-3y} \, dz \, dy \, dx + \int_0^\frac{2}{3} \int_0^{1-\frac{1}{2}y} \int_0^{3-3x-\frac{3}{2}y} \, dz \, dx \, dy = \frac{7}{8}
   $$
24. Find the mass of the solid determined by $25x^2 + 4y^2 \leq 9, z \geq 0$, and $z = x + 2$ if the density is $\rho(x, y, z) = x$.

Answer:

\[\int_{-\frac{3}{4}}^{\frac{3}{4}} \int_{\frac{2}{4}}^{\frac{3}{4}} \int_{0}^{x+2} (x) \, dz \, dy \, dx = \frac{81}{1000 \pi} \]

25. Find $\int_{0}^{1} \int_{0}^{35-5z} \int_{\frac{7-z}{4}}^{\frac{7-z}{4}} (7-z) \cos(y^2) \, dy \, dx \, dz$.

Answer:

You need to interchange the order of integration. $\int_{0}^{1} \int_{0}^{7-z} \int_{0}^{\frac{7-z}{4}} (7-z) \cos(y^2) \, dx \, dy \, dz = \frac{5}{4} \cos 36 - \frac{5}{4} \cos 49$

26. Find $\int_{0}^{2} \int_{0}^{12-3z} \int_{\frac{4-z}{4}}^{\frac{4-z}{4}} (4-z) \exp(y^2) \, dy \, dx \, dz$.

Answer:

You need to interchange the order of integration. $\int_{0}^{2} \int_{0}^{4-z} \int_{0}^{\frac{4-z}{4}} (4-z) \exp(y^2) \, dx \, dy \, dz = -\frac{3}{4} e^{4} - 9 + \frac{3}{4} e^{16}$

27. Find $\int_{0}^{2} \int_{0}^{25-5z} \int_{\frac{5-z}{4}}^{\frac{5-z}{4}} (5-z) \exp(x^2) \, dx \, dy \, dz$.

Answer:

You need to interchange the order of integration.

\[\int_{0}^{2} \int_{0}^{5-z} \int_{0}^{5x} (5-z) \exp(x^2) \, dx \, dy \, dz = -\frac{5}{4} e^{9} - 20 + \frac{5}{4} e^{25} \]

28. Find $\int_{0}^{10-2z} \int_{\frac{2x}{4}}^{5-z} \int_{0}^{\sin x} \frac{\sin x}{x} \, dx \, dy \, dz$.

Answer:

You need to interchange the order of integration.

\[\int_{0}^{1} \int_{0}^{5-z} \int_{0}^{2x} \frac{\sin x}{x} \, dy \, dx \, dz = -2 \sin 1 \cos 5 + 2 \cos 1 \sin 5 + 2 - 2 \sin 5 \]

29. Find $\int_{0}^{20} \int_{0}^{30} \int_{\frac{6-z}{4}}^{\frac{6-z}{4}} \sin x \, dx \, dz \, dy + \int_{0}^{30} \int_{0}^{6-z} \int_{\frac{6-z}{4}}^{\frac{6-z}{4}} \sin x \, dx \, dy \, dz$.

Answer:

You need to interchange the order of integration.

\[\int_{0}^{2} \int_{0}^{30-5z} \int_{\frac{6-z}{4}}^{\frac{6-z}{4}} \sin x \, dx \, dy \, dz = \int_{0}^{2} \int_{0}^{6-z} \int_{0}^{5z} \sin x \, dx \, dy \, dz = -5 \sin 2 \cos 6 + 5 \cos 2 \sin 6 + 10 - 5 \sin 6 \]
20.5 Exercises

1. Evaluate the integral $\int_{0}^{1} \int_{0}^{x} f(x, y, z) \, dy \, dx$

2. Find $\int_{0}^{3} \int_{0}^{2} \int_{0}^{x} f(x, y, z) \, dy \, dx$

3. Find $\int_{0}^{2} \int_{0}^{1} \int_{0}^{x} f(x, y, z) \, dy \, dx$

4. Evaluate the integral $\int_{0}^{3} \int_{0}^{2} \int_{0}^{x} f(x, y, z) \, dy \, dx$

5. Evaluate the integral $\int_{0}^{\pi} \int_{0}^{\pi} f(x, y, z) \, dy \, dx$

6. Evaluate the integral $\int_{0}^{\pi} \int_{0}^{4} f(x, y, z) \, dy \, dx$

7. Fill in the missing limits. $\int_{0}^{1} \int_{0}^{2} \int_{0}^{x} f(x, y, z) \, dy \, dx \, dz$

8. Find the volume of the region bounded by $z = 0$, $x = 0$, $y = 0$.

9. Find the volume of the region bounded by $z = 0$, $x = 0$, $y = 0$.

10. Find the mass of the bounded region, $R$, formed by the plane $x = 0$, $y = 0$, $z = 0$.

11. Find the mass of the bounded region, $R$, formed by the plane $x = 0$, $y = 0$, $z = 0$.

12. Here is an iterated integral: $\int_{0}^{1} \int_{0}^{1} \int_{0}^{x} f(x, y, z) \, dy \, dz \, dx$.

13. Find the volume of the bounded region determined by $2y + z = 3$, $x = 9 - y^2$, $y = 0$, $x = 0$.

14. Find the volume of the bounded region determined by $3y + 2z = 5$, $x = 9 - y^2$, $y = 0$, $x = 0$.

15. Find the volume of the bounded region determined by $5y + 2z = 3$, $x = 9 - y^2$, $y = 0$, $x = 0$.

16. Find the volume of the region bounded by $x^2 + y^2 = 25$, $z = x$, $z = 0$, and $x \geq 0$.

17. Find the volume of the region bounded by $x^2 + y^2 = 9$, $z = 3x$, $z = 0$, and $x \geq 0$.

18. Find the volume of the region determined by the intersection of the two cylinders, $x^2 + y^2 \leq 16$ and $y^2 + z^2 \leq 16$. 
19. Find the total mass of the bounded solid determined by \( z = 4 - x^2 - y^2 \) and \( x, y, z \geq 0 \) if the mass is given by \( \rho(x, y, z) = y \).

20. Find the total mass of the bounded solid determined by \( z = 9 - x^2 - y^2 \) and \( x, y, z \geq 0 \) if the mass is given by \( \rho(x, y, z) = z^2 \).

21. Find the volume of the region bounded by \( x^2 + y^2 = 4, z = 0, z = 5 - y \).

22. Find the volume of the bounded region determined by \( x \geq 0, y \geq 0, z \geq 0 \), and \( \frac{1}{3}x + \frac{3}{4}y + \frac{1}{2}z = 1 \), and \( \frac{1}{3}x + \frac{1}{4}y + \frac{1}{2}z = 1 \).

23. Find the volume of the bounded region determined by \( x \geq 0, y \geq 0, z \geq 0 \), and \( \frac{1}{3}x + \frac{1}{3}y + \frac{1}{3}z = 1 \), and \( \frac{1}{3}x + \frac{1}{3}y + \frac{1}{3}z = 1 \).

24. Find the mass of the solid determined by \( 16x^2 + 4y^2 \leq 9, z \geq 0 \), and \( z = x + 2 \) if the density is \( \rho(x, y, z) = z \).

25. Find \( \int_0^2 \int_0^{6-2z} \int_{\frac{3}{2}x}^{3-z} (3 - z) \cos(y^2) \, dy \, dx \, dz \).

26. Find \( \int_0^1 \int_0^{18-3z} \int_{\frac{3}{2}x}^{6-z} (6 - z) \exp(y^2) \, dy \, dx \, dz \).

27. Find \( \int_0^2 \int_0^{24-4z} \int_{\frac{3}{2}y}^{6-z} (6 - z) \exp(x^2) \, dx \, dy \, dz \).

28. Find \( \int_0^1 \int_0^{12-4z} \int_{\frac{3}{2}y}^{3-z} \frac{\sin x}{x} \, dx \, dy \, dz \).

29. Find \( \int_0^5 \int_0^{5-z} \frac{\sin x}{x} \, dx \, dz \, dy + \int_2^5 \int_0^{5-\frac{1}{2}y} \int_{\frac{3}{2}y}^{5-z} \frac{\sin x}{x} \, dx \, dz \, dy \). **Hint:** You might try doing it in the order, \( dy \, dx \, dz \).
THE RIEMANN INTEGRAL ON $\mathbb{R}^N$
The Integral In Other Coordinates

21.0.1 Outcomes

1. Represent a region in polar coordinates and use to evaluate integrals.
2. Represent a region in spherical or cylindrical coordinates and use to evaluate integrals.
3. Convert integrals in rectangular coordinates to integrals in polar coordinates and use to evaluate the integral.
4. Evaluate integrals in any coordinate system using the Jacobian.
5. Evaluate areas and volumes using another coordinate system.
6. Understand the transformation equations between spherical, polar and cylindrical coordinates and be able to change algebraic expressions from one system to another.
7. Use multiple integrals in an appropriate coordinate system to calculate the volume, mass, moments, center of gravity and moment of inertia.

21.1 Different Coordinates

As mentioned above, the fundamental concept of an integral is a sum of things of the form $f(x)\,dV$ where $dV$ is an “infinitesimal” chunk of volume located at the point, $x$. Up to now, this infinitesimal chunk of volume has had the form of a box with sides $dx_1, \cdots, dx_n$ and $dV = dx_1 \, dx_2 \cdots dx_n$ but its form is not important. It could just as well be an infinitesimal parallelepiped for example. In what follows, this is what it will be.

First recall the following fundamental definition on Page 190.

**Definition 21.1.1** Let $u_1, \cdots, u_p$ be vectors in $\mathbb{R}^k$. The parallelepiped determined by these vectors will be denoted by $P(u_1, \cdots, u_p)$ and it is defined as

$$P(u_1, \cdots, u_p) \equiv \left\{ \sum_{j=1}^{p} s_j u_j : s_j \in [0, 1] \right\}.$$

Now define the volume of this parallelepiped.

$\text{volume of } P(u_1, \cdots, u_p) \equiv \left(\det (u_i \cdot u_j)\right)^{1/2}$.
The dot product is used to determine this volume of a parallelepiped spanned by the given vectors and you should note that it is only the dot product that matters. Now consider spherical coordinates, \( \rho, \phi, \) and \( \theta \). Recall there is a relationship between these coordinates and rectangular coordinates given by

\[
x = \rho \sin \phi \cos \theta, \quad y = \rho \sin \phi \sin \theta, \quad z = \rho \cos \phi
\]

where \( \phi \in [0, \pi], \theta \in [0, 2\pi] \), and \( \rho > 0 \). Thus \((\rho, \phi, \theta)\) is a point in \( \mathbb{R}^3 \), more specifically in the set

\[
U = (0, \infty) \times [0, \pi] \times [0, 2\pi]
\]

and corresponding to such a \((\rho, \phi, \theta)\) \( \in U \) there exists a unique point, \((x, y, z)\) \( \in V \) where \( V \) consists of all points of \( \mathbb{R}^3 \) other than the origin, \((0, 0, 0)\). This \((x, y, z)\) determines a unique point in three dimensional space as mentioned earlier. Suppose at the point \((\rho_0, \phi_0, \theta_0)\) \( \in U \), there is an infinitesimal box having sides \( dp, d\phi, d\theta \). Then this little box would correspond to something in \( V \). What? Consider the mapping from \( U \) to \( V \) defined by

\[
x = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \rho \sin \phi \cos \theta \\ \rho \sin \phi \sin \theta \\ \rho \cos \phi \end{pmatrix} = f(\rho, \phi, \theta)
\]

which takes a point, \((\rho, \phi, \theta)\) in \( U \) and sends it to the point in \( V \) which is identified as \((x, y, z)^T \equiv x\). What happens to a point of the infinitesimal box, located at \((\rho_0, \phi_0, \theta_0)\)? Such a point is of the form

\[
(\rho_0 + s_1 dp, \phi_0 + s_2 d\phi, \theta_0 + s_3 d\theta),
\]

where \( s_i \geq 0 \) and \( \sum_i s_i \leq 1 \). Also, from the definition of the derivative,

\[
f(\rho_0 + s_1 dp, \phi_0 + s_2 d\phi, \theta_0 + s_3 d\theta) - f(\rho_0, \phi_0, \theta_0) =
\]

\[
Df(\rho_0, \phi_0, \theta_0) \begin{pmatrix} s_1 dp \\ s_2 d\phi \\ s_3 d\theta \end{pmatrix} + o \begin{pmatrix} s_1 dp \\ s_2 d\phi \\ s_3 d\theta \end{pmatrix}
\]

where the last term may be taken equal to \( \mathbf{0} \) because the vector, \((s_1 dp, s_2 d\phi, s_3 d\theta)^T\) is infinitesimal meaning nothing precise but conveying the idea that it is surpassingly small. Therefore, a point of this infinitesimal box is sent to the vector,

\[
\begin{pmatrix}
\frac{\partial x(\rho_0, \phi_0, \theta_0)}{\partial \rho}, & \frac{\partial x(\rho_0, \phi_0, \theta_0)}{\partial \phi}, & \frac{\partial x(\rho_0, \phi_0, \theta_0)}{\partial \theta}
\end{pmatrix}
\]

\[
\begin{pmatrix}
s_1 dp \\ s_2 d\phi \\ s_3 d\theta
\end{pmatrix}
\]

a point of the infinitesimal parallelepiped determined by the vectors

\[
\left\{ \frac{\partial x(\rho_0, \phi_0, \theta_0)}{\partial \rho} dp, \frac{\partial x(\rho_0, \phi_0, \theta_0)}{\partial \phi} d\phi, \frac{\partial x(\rho_0, \phi_0, \theta_0)}{\partial \theta} d\theta \right\}.
\]

The situation is no different for general coordinate systems. In general, \( \mathbf{x} = f(\mathbf{u}) \) where \( \mathbf{u} \in U \), a subset of \( \mathbb{R}^n \) and \( \mathbf{x} \) is a point in \( V \), a subset of \( n \) dimensional space. Thus, letting the Cartesian coordinates of \( \mathbf{x} \) be given by \( \mathbf{x} = (x_1, \ldots, x_n)^T \), each \( x_i \) being a function of
\( u \), an infinitesimal box located at \( u_0 \) corresponds to an infinitesimal parallelepiped located at \( f(u_0) \) which is determined by the \( n \) vectors \( \left\{ \frac{\partial x(u_0)}{\partial u_i} \right\}_{i=1}^n \). From Definition 21.1.1, the volume of this infinitesimal parallelepiped located at \( f(u_0) \) is given by
\[
\det \left( \frac{\partial x(u_0)}{\partial u_i} \cdot \frac{\partial x(u_0)}{\partial u_j} \right)^{1/2}
\]
(21.3)
in which there is no sum on the repeated index. Now in general if there are \( n \) vectors in \( \mathbb{R}^n \), \( \{v_1, \cdots, v_n\} \),
\[
\det (v_i \cdot v_j)^{1/2} = |\det (v_1, \cdots, v_n)|
\]
(21.4)
where this last matrix is the \( n \times n \) matrix which has the \( i^{th} \) column equal to \( v_i \). The reason for this is that the matrix whose \( ij^{th} \) entry is \( v_i \cdot v_j \) is just the product of the two matrices,
\[
\begin{pmatrix}
    v_1^T \\
    \vdots \\
    v_n^T
\end{pmatrix}
(v_1, \cdots, v_n)
\]
where the first on the left is the matrix having the \( i^{th} \) row equal to \( v_i^T \) while the matrix on the right is just the matrix having the \( i^{th} \) column equal to \( v_i \). Therefore, since the determinant of a matrix equals the determinant of its transpose,
\[
\det (v_i \cdot v_j) = \det \left( \begin{pmatrix}
    v_1^T \\
    \vdots \\
    v_n^T
\end{pmatrix}
(v_1, \cdots, v_n)
\right)
= \det (v_1, \cdots, v_n)^2
\]
and so taking square roots yields (21.4). Therefore, from the properties of determinants, (21.3) equals
\[
\left| \det \left( \frac{\partial x(u_0)}{\partial u_1}, \cdots, \frac{\partial x(u_0)}{\partial u_n} \right) \right| =
\left| \det \left( \frac{\partial x(u_0)}{\partial u_1}, \cdots, \frac{\partial x(u_0)}{\partial u_n} \right) \right| du_1 \cdots du_n
\]
and this is the infinitesimal chunk of volume corresponding to the point \( f(u_0) \) in \( V \).

**Definition 21.1.2** Let \( x = f(u) \) be as described above. Then the symbol \( \frac{\partial (x_1, \cdots, x_n)}{\partial (u_1, \cdots, u_n)} \), called the Jacobian determinant, is defined by
\[
\det \left( \frac{\partial x(u_0)}{\partial u_1}, \cdots, \frac{\partial x(u_0)}{\partial u_n} \right) = \frac{\partial (x_1, \cdots, x_n)}{\partial (u_1, \cdots, u_n)}.
\]

Also, the symbol, \( \left| \frac{\partial (x_1, \cdots, x_n)}{\partial (u_1, \cdots, u_n)} \right| du_1 \cdots du_n \) is called the volume element.

This has given motivation for the following fundamental procedure often called the change of variables formula which holds under fairly general conditions.
Procedure 21.1.3 Suppose $U$ is a subset of $\mathbb{R}^n$ and suppose $f : U \rightarrow V$ is a $C^1$ function which is one to one.\(^1\) Then if $h : V \rightarrow \mathbb{R}$,

$$\int_U h(f(u)) \left| \frac{\partial (x_1, \cdots, x_n)}{\partial (u_1, \cdots, u_n)} \right| dV = \int_V h(x) \, dV.$$ \(\square\)

Now return to Spherical coordinates. In this case, it is necessary to find the absolute value of

$$\det \left( \frac{\partial x}{\partial \rho}, \frac{\partial x}{\partial \phi}, \frac{\partial x}{\partial \theta} \right)$$

which equals

$$\rho^2 \sin \phi$$

which is positive because $\phi \in [0, \pi]$.

Example 21.1.4 Find the volume of a ball, $B_R$ of radius $R$.

In this case, $U = (0, R] \times [0, \pi] \times [0, 2\pi)$ and use spherical coordinates. Then (21.2) yields a set in $\mathbb{R}^3$ which clearly differs from the ball of radius $R$ only by a set having volume equal to zero. It leaves out the point at the origin is all. Therefore, the volume of the ball is

$$\int_{B_R} 1 \, dV = \int_U \rho^2 \sin \phi \, dV$$

$$= \int_0^R \int_0^\pi \int_0^{2\pi} \rho^2 \sin \phi \, d\phi \, d\theta \, d\rho = \frac{4}{3} R^3 \pi.$$ \(\square\)

The reason this was effortless, is that the ball, $B_R$ is realized as a box in terms of the spherical coordinates. Remember what was pointed out earlier about setting up iterated integrals over boxes.

Example 21.1.5 Find the volume element for cylindrical coordinates.

In cylindrical coordinates,

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r \cos \theta \\ r \sin \theta \\ z \end{pmatrix}$$

Therefore, the Jacobian determinant is

$$\det \begin{pmatrix} \cos \theta & -r \sin \theta & 0 \\ \sin \theta & r \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix} = r.$$ \(\square\)

It follows the volume element in cylindrical coordinates is $r \, d\theta \, dr \, dz$.

\(^1\)This will cause non overlapping infinitesimal boxes in $U$ to be mapped to non overlapping infinitesimal parallelepipeds in $V$.

Also, in the context of the Riemann integral we should say more about the sets, $U$ and $V$ in any case the function, $h$. These conditions are mainly technical however, and since a mathematically respectable treatment will not be attempted for this theorem, I think it best to give a memorable version of it which is essentially correct in all examples of interest.
Example 21.1.6 This example uses spherical coordinates to verify an important conclusion about gravitational force. Let the hollow sphere, $H$ be defined by $a^2 \leq x^2 + y^2 + z^2 \leq b^2$ and suppose this hollow sphere has constant density taken to equal 1. Now place a unit mass at the point $(0,0,z_0)$ where $|z_0| \in [a,b]$. Show the force of gravity acting on this unit mass is 
\[
\alpha G \int \int_V \frac{(z-z_0)}{[x^2+y^2+(z-z_0)^2]^{3/2}} \, dV \k \quad \text{and then show that if } |z_0| > b \text{ then the force of gravity acting on this point mass is the same as if the entire mass of the hollow sphere were placed at the origin, while if } |z_0| < a, \text{ the total force acting on the point mass from gravity equals zero}. \quad \text{Here } G \text{ is the gravitation constant and } \alpha \text{ is the density. In particular, this shows that the force a planet exerts on an object is as though the entire mass of the planet were situated at its center.}
\]

Without loss of generality, assume $z_0 > 0$. Let $dV$ be a little chunk of material located at the point $(x,y,z)$ of $H$ the hollow sphere. Then according to Newton’s law of gravity, the force this small chunk of material exerts on the given point mass equals

\[
\frac{x\mathbf{i} + y\mathbf{j} + (z-z_0) \mathbf{k}}{|x\mathbf{i} + y\mathbf{j} + (z-z_0) \mathbf{k}|} \frac{1}{(x^2+y^2+(z-z_0)^2)^{3/2}} \alpha \, dV = \frac{(x\mathbf{i} + y\mathbf{j} + (z-z_0) \mathbf{k})}{(x^2+y^2+(z-z_0)^2)^{3/2}} \alpha \, dV.
\]

Therefore, the total force is

\[
\int \int \int_H (x\mathbf{i} + y\mathbf{j} + (z-z_0) \mathbf{k}) \frac{1}{(x^2+y^2+(z-z_0)^2)^{3/2}} \alpha \, dV.
\]

By the symmetry of the sphere, the i and j components will cancel out when the integral is taken. This is because there is the same amount of stuff for negative $x$ and $y$ as there is for positive $x$ and $y$. Hence what remains is

\[
\alpha G \mathbf{k} \int \int \int_H \frac{(z-z_0)}{[x^2+y^2+(z-z_0)^2]^{3/2}} \, dV
\]

as claimed. Now for the interesting part, the integral is evaluated. In spherical coordinates this integral is,

\[
\int_0^{2\pi} \int_a^b \int_0^{\pi} \frac{(\rho \cos \phi - z_0) \rho^2 \sin \phi}{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)^{3/2}} \, d\phi \, d\rho \, d\theta.
\]

Rewrite the inside integral and use integration by parts to obtain this inside integral equals

\[
\frac{1}{2z_0} \int_0^{\pi} \frac{\rho^2 \cos \phi - \rho z_0}{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)^{3/2}} \, d\phi = 
\frac{1}{2z_0} \left( \frac{-\rho^2 - \rho z_0}{\sqrt{(\rho^2 + z_0^2 + 2\rho z_0)}} + 2 \frac{\rho^2 - \rho z_0}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0)}} - \int_0^{\pi} 2\rho^2 \frac{\sin \phi}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)}} \, d\phi \right). \quad (21.6)
\]

There are some cases to consider here.

\footnote{This was shown by Newton in 1685 and allowed him to assert his law of gravitation applied to the planets as though they were point masses. It was a major accomplishment.}
First suppose \( z_0 < a \) so the point is on the inside of the hollow sphere and it is always the case that \( \rho > z_0 \). Then in this case, the two first terms reduce to
\[
\frac{2\rho (\rho + z_0)}{\sqrt{(\rho + z_0)^2}} + \frac{2\rho (\rho - z_0)}{\sqrt{(\rho - z_0)^2}} = \frac{2\rho (\rho + z_0)}{\rho} + \frac{2\rho (\rho - z_0)}{\rho - z_0} = 4\rho
\]
and so the expression in (21.6) equals
\[
\frac{1}{2z_0} \left( 4\rho - \int_0^\pi 2\rho^2 \frac{\sin \phi}{\sqrt{\rho^2 + z_0^2 - 2\rho z_0 \cos \phi}} \, d\phi \right)
\]
\[
= \frac{1}{2z_0} \left( 4\rho - \frac{1}{z_0} \int_0^\pi \rho \frac{2\rho z_0 \sin \phi}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)}} \, d\phi \right)
\]
\[
= \frac{1}{2z_0} \left( 4\rho - \frac{2\rho}{z_0} \left( \rho^2 + z_0^2 - 2\rho z_0 \cos \phi \right)^{1/2} \right)|_0^\pi
\]
\[
= \frac{1}{2z_0} \left( 4\rho - \frac{2\rho}{z_0} \left[ (\rho + z_0) - (\rho - z_0) \right] \right) = 0.
\]
Therefore, in this case the inner integral of (21.5) equals zero and so the original integral will also be zero.

The other case is when \( z_0 > b \) and so it is always the case that \( z_0 > \rho \). In this case the first two terms of (21.6) are
\[
\frac{2\rho (\rho + z_0)}{\sqrt{(\rho + z_0)^2}} + \frac{2\rho (\rho - z_0)}{\sqrt{(\rho - z_0)^2}} = \frac{2\rho (\rho + z_0)}{(\rho + z_0)} + \frac{2\rho (\rho - z_0)}{z_0 - \rho} = 0.
\]

Therefore in this case, (21.6) equals
\[
\frac{1}{2z_0} \left( - \int_0^\pi 2\rho^2 \frac{\sin \phi}{\sqrt{\rho^2 + z_0^2 - 2\rho z_0 \cos \phi}} \, d\phi \right)
\]
\[
= -\frac{\rho}{2z_0} \left( \int_0^\pi \frac{2\rho z_0 \sin \phi}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)}} \, d\phi \right)
\]
which equals
\[
-\frac{\rho}{z_0} \left( \left( \rho^2 + z_0^2 - 2\rho z_0 \cos \phi \right)^{1/2} \right)|_0^\pi
\]
\[
= -\frac{\rho}{z_0} \left[ (\rho + z_0) - (\rho - z_0) \right] = -\frac{2\rho^2}{z_0}.
\]

Thus the inner integral of (21.5) reduces to the above simple expression. Therefore, (21.5) equals
\[
\int_0^{2\pi} \int_a^b \left( -\frac{2}{z_0}\rho^2 \right) \, d\rho \, d\theta = -\frac{4}{3} \pi \frac{b^3 - a^3}{z_0}
\]
and so
\[
\alpha Gk \int \int_H \frac{(z - z_0)}{\left( x^2 + y^2 + (z - z_0)^2 \right)^{3/2}} \, dV = \alpha Gk \left( -\frac{4}{3} \pi \frac{b^3 - a^3}{z_0} \right) = -kG \frac{\text{total mass}}{z_0^2}.
\]
21.2 Exercises With Answers

1. Find the area of the bounded region, \( R \), determined by \( 3x + 3y = 1 \), \( 3x + 3y = 8 \), \( y = 3x \), and \( y = 4x \).

Answer:

Let \( u = \frac{y}{3} \), \( v = 3x + 3y \). Then solving these equations for \( x \) and \( y \) yields

\[
\begin{cases} 
  x = \frac{1}{3} \frac{v}{1+u} \cdot y = \frac{1}{3} \frac{v}{1+u} 
\end{cases}.
\]

Now

\[
\frac{\partial (x, y)}{\partial (u, v)} = \det \begin{pmatrix} -\frac{1}{3} \frac{v}{(1+u)^2} & \frac{3+3u}{3+3u} \\ \frac{3+3u}{3+3u} & \frac{v}{(1+u)} \end{pmatrix} = -\frac{1}{9} \frac{v}{(1+u)^2}.
\]

Also, \( u \in [3, 4] \) while \( v \in [1, 8] \). Therefore,

\[
\int_R dV = \int_3^4 \int_1^8 \left| -\frac{1}{9} \frac{v}{(1+u)^2} \right| dv du = \int_3^4 \int_1^8 \frac{1}{9} \frac{v}{(1+u)^2} dv du = \frac{7}{40}.
\]

2. Find the area of the bounded region, \( R \), determined by \( 5x + y = 1 \), \( 5x + y = 9 \), \( y = 2x \), and \( y = 5x \).

Answer:

Let \( u = \frac{y}{5} \), \( v = 5x + y \). Then solving these equations for \( x \) and \( y \) yields

\[
\begin{cases} 
  x = \frac{v}{5+u} \cdot y = \frac{v}{5+u} 
\end{cases}.
\]

Now

\[
\frac{\partial (x, y)}{\partial (u, v)} = \det \begin{pmatrix} -\frac{v}{(5+u)^2} & \frac{1}{5+u} \\ \frac{1}{5+u} & \frac{v}{5+u} \end{pmatrix} = -\frac{v}{(5+u)^2}.
\]

Also, \( u \in [2, 5] \) while \( v \in [1, 9] \). Therefore,

\[
\int_R dV = \int_2^5 \int_1^9 \left| -\frac{v}{(5+u)^2} \right| dv du = \int_2^5 \int_1^9 \frac{v}{(5+u)^2} dv du = \frac{12}{7}.
\]

3. A solid, \( R \) is determined by \( 5x + 3y = 4 \), \( 5x + 3y = 9 \), \( y = 2x \), and \( y = 5x \) and the density is \( \rho = x \). Find the total mass of \( R \).

Answer:

Let \( u = \frac{y}{3} \), \( v = 5x + 3y \). Then solving these equations for \( x \) and \( y \) yields

\[
\begin{cases} 
  x = \frac{v}{5+3u} \cdot y = \frac{v}{5+3u} 
\end{cases}.
\]

Now
\[
\frac{\partial (x, y)}{\partial (u, v)} = \det \left( \begin{array}{cc}
-\frac{3}{(5+3u)^2} & \frac{1}{5+3u} \\
\frac{1}{5+3u} & \frac{1}{5+3u^2}
\end{array} \right) = -\frac{v}{(5+3u)^2}.
\]

Also, \( u \in [2, 5] \) while \( v \in [4, 9] \). Therefore,

\[
\int_R \rho \, dV = \int_2^5 \int_4^9 \frac{v}{5+3u} \left| -\frac{v}{(5+3u)^2} \right| \, dv \, du = -4 \ln 2 + 4 \ln 3.
\]
6. Find the volume of the region, \( E \), bounded by the ellipsoid, \( \frac{1}{4} x^2 + \frac{1}{9} y^2 + \frac{1}{49} z^2 = 1 \).

Answer:
Let \( u = \frac{1}{2} x, v = \frac{1}{3} y, w = \frac{1}{7} z \). Then \((u, v, w)\) is a point in the unit ball, \( B \). Therefore,
\[
\int_B \frac{\partial (x, y, z)}{\partial (u, v, w)} \, dV = \int_E dV.
\]
But \( \frac{\partial (x, y, z)}{\partial (u, v, w)} = 42 \) and so the answer is
\[
(volume \ of \ B) \times 42 = \frac{4}{3} \pi 42 = 56\pi.
\]

7. Here are three vectors. \((4, 1, 4)^T, (5, 0, 4)^T, \) and \((3, 1, 5)^T\). These vectors determine a parallelepiped, \( R \), which is occupied by a solid having density \( \rho = x \). Find the mass of this solid.

Answer:
Let \[
\begin{pmatrix} 4 & 5 & 3 \\ 1 & 0 & 1 \\ 4 & 4 & 5 \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}.
\]
Then this maps the unit cube,
\[
Q \equiv [0, 1] \times [0, 1] \times [0, 1]
\]
on to \( R \) and
\[
\frac{\partial (x, y, z)}{\partial (u, v, w)} = \left| \det \begin{pmatrix} 4 & 5 & 3 \\ 1 & 0 & 1 \\ 4 & 4 & 5 \end{pmatrix} \right| = | -9 | = 9
\]
so the mass is
\[
\int_R x \, dV = \int_Q (4u + 5v + 3w) (9) \, dV
\]
\[
= \int_0^1 \int_0^1 \int_0^1 (4u + 5v + 3w) (9) \, du \, dv \, dw = 54
\]

8. Here are three vectors. \((3, 2, 6)^T, (4, 1, 6)^T, \) and \((2, 2, 7)^T\). These vectors determine a parallelepiped, \( R \), which is occupied by a solid having density \( \rho = y \). Find the mass of this solid.

Answer:
Let \[
\begin{pmatrix} 3 & 4 & 2 \\ 2 & 1 & 2 \\ 6 & 6 & 7 \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}.
\]
Then this maps the unit cube,
\[
Q \equiv [0, 1] \times [0, 1] \times [0, 1]
\]
on to \( R \) and
\[
\frac{\partial (x, y, z)}{\partial (u, v, w)} = \left| \det \begin{pmatrix} 3 & 4 & 2 \\ 2 & 1 & 2 \\ 6 & 6 & 7 \end{pmatrix} \right| = | -11 | = 11
\]
and so the mass is
\[
\int_R x \, dV = \int_Q (2u + v + 2w) \, dV
\]
\[
= \int_0^1 \int_0^1 \int_0^1 (2u + v + 2w) \, du \, dv \, dw = \frac{55}{2}.
\]

9. Here are three vectors. \((2,2,4)^T, (3,1,4)^T, \) and \((1,2,5)^T\). These vectors determine a parallelepiped, \(R\), which is occupied by a solid having density \(\rho = y + x\). Find the mass of this solid.

Answer:
Let \[
\begin{pmatrix}
2 & 3 & 1 \\
2 & 1 & 2 \\
4 & 4 & 5
\end{pmatrix}
\begin{pmatrix}
u \\
v \\
w
\end{pmatrix} = \begin{pmatrix}x \\
y \\
z
\end{pmatrix}.
\]
Then this maps the unit cube,
\[
Q \equiv [0,1] \times [0,1] \times [0,1]
\]
on to \(R\) and
\[
\frac{\partial (x,y,z)}{\partial (u,v,w)} = \left| \det \begin{pmatrix}
2 & 3 & 1 \\
2 & 1 & 2 \\
4 & 4 & 5
\end{pmatrix} \right| = |-8| = 8
\]
and so the mass is \(2u + 3v + w\)
\[
\int_R x \, dV = \int_Q (4u + 4v + 3w) \, dV
\]
\[
= \int_0^1 \int_0^1 \int_0^1 (4u + 4v + 3w) \, du \, dv \, dw = 44.
\]

10. Let \(D = \{(x,y) : x^2 + y^2 \leq 25\}\). Find \(\int_D e^{36x^2 + 36y^2} \, dx \, dy\).

Answer:
This is easy in polar coordinates. \(x = r \cos \theta, y = r \sin \theta\). Thus \(\frac{\partial (x,y)}{\partial (r,\theta)} = r\) and in terms of these new coordinates, the disk, \(D\), is the rectangle,
\[
R = \{(r,\theta) : 0 \leq r \leq 5, 0 \leq \theta \leq 2\pi\}.
\]
Therefore,
\[
\int_D e^{36x^2 + 36y^2} \, dV = \int_R e^{36r^2} \, r \, dV =
\]
\[
\int_0^5 \int_0^{2\pi} e^{36r^2} \, r \, d\theta \, dr = \frac{1}{36} \pi (e^{900} - 1).
\]
Note you wouldn’t get very far without changing the variables in this.
11. Let \( D = \{ (x, y) : x^2 + y^2 \leq 9 \} \). Find \( \int_D \cos (36x^2 + 36y^2) \, dx \, dy \).

Answer:
This is easy in polar coordinates. \( x = r \cos \theta, y = r \sin \theta \). Thus \( \frac{\partial (x, y)}{\partial (r, \theta)} = r \) and in terms of these new coordinates, the disk, \( D \), is the rectangle,

\[ R = \{ (r, \theta) \in [0, 3] \times [0, 2\pi] \} . \]

Therefore,

\[
\int_D \cos (36x^2 + 36y^2) \, dV = \int_R \cos (36r^2) \, r \, dV = \\
\int_0^3 \int_0^{2\pi} \cos (36r^2) \, r \, d\theta \, dr = \frac{1}{36} (\sin 324) \pi .
\]

12. The ice cream in a sugar cone is described in spherical coordinates by \( \rho \in [0, 8], \phi \in [0, \frac{\pi}{4}], \theta \in [0, 2\pi] \). If the units are in centimeters, find the total volume in cubic centimeters of this ice cream.

Answer:
Remember that in spherical coordinates, the volume element is \( \rho^2 \sin \phi \, dV \) and so the total volume of this is

\[
\frac{1}{3} \int_0^8 \int_0^{\frac{\pi}{4}} \int_0^{2\pi} \rho^2 \sin \phi \, d\phi \, d\theta \, d\rho = \frac{512}{3} \sqrt{2} \pi + \frac{1024}{3} \pi .
\]

13. Find the volume between \( z = 5 - x^2 - y^2 \) and \( z = \sqrt{x^2 + y^2} \).

Answer:
Use cylindrical coordinates. In terms of these coordinates the shape is

\( h - r^2 \geq z \geq r, r \in \left[ 0, \frac{1}{2} \sqrt{21} - \frac{1}{2} \right], \theta \in [0, 2\pi] \).

Also, \( \frac{\partial (x, y, z)}{\partial (r, \theta, z)} = r \). Therefore, the volume is

\[
\int_0^{\frac{1}{2} \sqrt{21} - \frac{1}{2}} \int_0^{2\pi} \int_0^{\sqrt{21 - r^2}} r \, dz \, d\theta \, dr = \frac{39}{4} \pi + \frac{1}{4} \pi \sqrt{21} .
\]

14. A ball of radius 12 is placed in a drill press and a hole of radius 4 is drilled out with the center of the hole a diameter of the ball. What is the volume of the material which remains?

Answer:
You know the formula for the volume of a sphere and so if you find out how much stuff is taken away, then it will be easy to find what is left. To find the volume of what is removed, it is easiest to use cylindrical coordinates. This volume is

\[
\int_0^4 \int_0^{2\pi} \int_{-(144-r^2)}^{(144-r^2)} r \, dz \, d\theta \, dr = -\frac{4096}{3} \sqrt{2} \pi + 2304 \pi .
\]

Therefore, the volume of what remains is \( \frac{4}{3} \pi (12)^3 \) minus the above. Thus the volume of what remains is

\[
\frac{4096}{3} \sqrt{2} \pi .
\]
15. A ball of radius 11 has density equal to $\sqrt{x^2 + y^2 + z^2}$ in rectangular coordinates. The top of this ball is sliced off by a plane of the form $z = 1$. What is the mass of what remains?

Answer:

$$\int_0^{2\pi} \int_0^{\arcsin\left(\frac{2}{\sqrt{30}}\right)} \rho^3 \sin \phi \, d\rho \, d\phi \, d\theta + \int_0^{2\pi} \int_{\arcsin\left(\frac{2}{\sqrt{30}}\right)}^{\pi/2} \rho^3 \sin \phi \, d\rho \, d\phi \, d\theta$$

$$= \frac{24623}{3 \pi}$$

16. Find $\int_S \frac{y}{x} \, dV$ where $S$ is described in polar coordinates as $1 \leq r \leq 2$ and $0 \leq \theta \leq \pi/4$.

Answer:

Use $x = r \cos \theta$ and $y = r \sin \theta$. Then the integral in polar coordinates is

$$\int_0^{\pi/4} \int_1^2 (r \tan \theta) \, dr \, d\theta = \frac{3}{4} \ln 2.$$

17. Find $\int_S \left(\left(\frac{x}{y}\right)^2 + 1\right) \, dV$ where $S$ is given in polar coordinates as $1 \leq r \leq 2$ and $0 \leq \theta \leq \frac{1}{4} \pi$.

Answer:

Use $x = r \cos \theta$ and $y = r \sin \theta$. Then the integral in polar coordinates is

$$\int_0^{\pi/4} \int_1^2 \left(1 + \tan^2 \theta\right) r \, dr \, d\theta.$$

18. Use polar coordinates to evaluate the following integral. Here $S$ is given in terms of the polar coordinates. $\int_S \sin \left(4x^2 + 4y^2\right) \, dV$ where $r \leq 2$ and $0 \leq \theta \leq \frac{1}{4} \pi$.

Answer:

$$\int_0^{\pi/4} \int_1^2 \sin \left(4r^2\right) r \, dr \, d\theta.$$

19. Find $\int_S e^{2x^2 + 2y^2} \, dV$ where $S$ is given in terms of the polar coordinates, $r \leq 2$ and $0 \leq \theta \leq \frac{1}{4} \pi$.

Answer:

The integral is

$$\int_0^{\pi/4} \int_0^2 e^{2r^2} \, dr \, d\theta = \frac{1}{12} \pi \left(e^8 - 1\right).$$

20. Compute the volume of a sphere of radius $R$ using cylindrical coordinates.

Answer:

Using cylindrical coordinates, the integral is $\int_0^{2\pi} \int_0^R e^{\sqrt{R^2 - r^2}} r \, dz \, d\theta = \frac{4}{3} \pi R^3.$
21.3 Exercises

1. Find the area of the bounded region, $R$, determined by $5x + y = 2, 5x + y = 8, y = 2x$, and $y = 6x$.

2. Find the area of the bounded region, $R$, determined by $y + 2x = 6, y + 2x = 10, y = 3x$, and $y = 4x$.

3. A solid, $R$ is determined by $3x + y = 2, 3x + y = 4, y = 2x$, and $y = 6x$ and the density is $\rho = x$. Find the total mass of $R$.

4. A solid, $R$ is determined by $4x + 2y = 5, 4x + 2y = 6, y = 5x$, and $y = 7x$ and the density is $\rho = y$. Find the total mass of $R$.

5. A solid, $R$ is determined by $3x + y = 3, 3x + y = 10, y = 3x$, and $y = 5x$ and the density is $\rho = y^{-1}$. Find the total mass of $R$.

6. Find the volume of the region, $E$, bounded by the ellipsoid, $\frac{1}{4}x^2 + y^2 + z^2 = 1$.

7. Here are three vectors. $(4,1,2)^T, (5,0,2)^T, (3,1,3)^T$. These vectors determine a parallelepiped, $R$, which is occupied by a solid having density $\rho = x$. Find the mass of this solid.

8. Here are three vectors. $(5,1,6)^T, (6,0,6)^T, (4,1,7)^T$. These vectors determine a parallelepiped, $R$, which is occupied by a solid having density $\rho = y$. Find the mass of this solid.

9. Here are three vectors. $(5,2,9)^T, (6,1,9)^T, (4,2,10)^T$. These vectors determine a parallelepiped, $R$, which is occupied by a solid having density $\rho = y + x$. Find the mass of this solid.

10. Let $D = \{(x,y) : x^2 + y^2 \leq 25]\}. \text{ Find } \int_D e^{25x^2 + 25y^2} \, dx \, dy$.

11. Let $D = \{(x,y) : x^2 + y^2 \leq 16\}. \text{ Find } \int_D \cos(9x^2 + 9y^2) \, dx \, dy$.

12. The ice cream in a sugar cone is described in spherical coordinates by $\rho \in [0,10], \phi \in [0,\frac{\pi}{2}], \theta \in [0,2\pi]$. If the units are in centimeters, find the total volume in cubic centimeters of this ice cream.

13. Find the volume between $z = 5 - x^2 - y^2$ and $z = 2\sqrt{x^2 + y^2}$.

14. A ball of radius 3 is placed in a drill press and a hole of radius 2 is drilled out with the center of the hole a diameter of the ball. What is the volume of the material which remains?

15. A ball of radius 9 has density equal to $\sqrt{x^2 + y^2 + z^2}$ in rectangular coordinates. The top of this ball is sliced off by a plane of the form $z = 2$. What is the mass of what remains?

16. Find $\int_S \frac{y}{2} \, dV$ where $S$ is described in polar coordinates as $1 \leq r \leq 2$ and $0 \leq \theta \leq \pi/4$.

17. Find $\int_S \left(\frac{y}{r}\right)^2 + 1 \, dV$ where $S$ is given in polar coordinates as $1 \leq r \leq 2$ and $0 \leq \theta \leq \frac{3}{2}\pi$.

18. Use polar coordinates to evaluate the following integral. Here $S$ is given in terms of the polar coordinates. $\int_S \sin(2x^2 + 2y^2) \, dV$ where $r \leq 2$ and $0 \leq \theta \leq \frac{3}{2}\pi$. 


19. Find \[ \int \int_S e^{2x^2+2y^2} \, dV \] where \( S \) is given in terms of the polar coordinates, \( r \leq 2 \) and \( 0 \leq \theta \leq \pi \).

20. Compute the volume of a sphere of radius \( R \) using cylindrical coordinates.

21. In Example 21.1.6 on Page 385 check out all the details by working the integrals to be sure the steps are right.

22. What if the hollow sphere in Example 21.1.6 were in two dimensions and everything, including Newton’s law still held? Would similar conclusions hold? Explain.

23. Fill in all details for the following argument that \( \int_0^\infty e^{-x^2} \, dx = \frac{1}{2} \sqrt{\pi} \). Let \( I = \int_0^\infty e^{-x^2} \, dx \).

Then

\[
I^2 = \int_0^\infty \int_0^\infty e^{-(x^2+y^2)} \, dx \, dy = \int_0^{\pi/2} \int_0^\infty re^{-r^2} \, dr \, d\theta = \frac{1}{4} \pi
\]

from which the result follows.

24. Show using Problem 23 \( \Gamma \left( \frac{1}{2} \right) = \sqrt{\pi} \).

25. Let \( p, q > 0 \) and define \( B(p, q) = \int_0^1 x^{p-1} (1-x)^{q-1} \). Show \( \Gamma(p) \Gamma(q) = B(p, q) \Gamma(p + q) \).

**Hint:** It is fairly routine if you start with the left side and proceed to change variables.

### 21.4 The Moment Of Inertia

In order to appreciate the importance of this concept, it is necessary to discuss its physical significance.

#### 21.4.1 The Spinning Top

To begin with consider a spinning top as illustrated in the following picture.
For the purpose of this discussion, consider the top as a large number of point masses, \( m_i \), located at the positions, \( r_i(t) \) for \( i = 1, 2, \ldots, N \) and these masses are symmetrically arranged relative to the axis of the top. As the top spins, the axis of symmetry is observed to move around the \( z \) axis. This is called precession and you will see it occur whenever you spin a top. What is the speed of this precession? In other words, what is \( \theta' \)? The following discussion follows one given in Sears and Zemansky [25].

Imagine a coordinate system which is fixed relative to the moving top. Thus in this coordinate system the points of the top are fixed. Let the standard unit vectors of the coordinate system moving with the top be denoted by \( \mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t) \). From Theorem 13.4.2 on Page 241, there exists an angular velocity vector \( \Omega(t) \) such that if \( u(t) \) is the position vector of a point fixed in the top, \( u(t) = u_1 \mathbf{i}(t) + u_2 \mathbf{j}(t) + u_3 \mathbf{k}(t) \),

\[
\mathbf{u}'(t) = \Omega(t) \times \mathbf{u}(t).
\]

The vector \( \Omega_a \) shown in the picture is the vector for which

\[
r_i'(t) \equiv \Omega_a \times r_i(t)
\]

is the velocity of the \( i^{th} \) point mass due to rotation about the axis of the top. Thus \( \Omega(t) = \Omega_a(t) + \Omega_p(t) \) and it is assumed \( \Omega_p(t) \) is very small relative to \( \Omega_a \). In other words, it is assumed the axis of the top moves very slowly relative to the speed of the points in the top which are spinning very fast around the axis of the top. The angular momentum, \( \mathbf{L} \) is defined by

\[
\mathbf{L} \equiv \sum_{i=1}^{N} r_i \times m_i v_i
\]  \hspace{1cm} (21.7)

where \( v_i \) equals the velocity of the \( i^{th} \) point mass. Thus \( v_i = \Omega(t) \times r_i \) and from the above
assumption, \( \mathbf{v}_i \) may be taken equal to \( \Omega_a \times \mathbf{r}_i \). Therefore, \( \mathbf{L} \) is essentially given by

\[
\mathbf{L} \equiv \sum_{i=1}^{N} m_i \mathbf{r}_i \times (\Omega_a \times \mathbf{r}_i)
\]

\[
= \sum_{i=1}^{N} m_i \left( |\mathbf{r}_i|^2 \Omega_a - (\mathbf{r}_i \cdot \Omega_a) \mathbf{r}_i \right).
\]

By symmetry of the top, this last expression equals a multiple of \( \Omega_a \). Thus \( \mathbf{L} \) is parallel to \( \Omega_a \). Also,

\[
\mathbf{L} \cdot \Omega_a = \sum_{i=1}^{N} m_i \Omega_a \cdot \mathbf{r}_i \times (\Omega_a \times \mathbf{r}_i)
\]

\[
= \sum_{i=1}^{N} m_i (\Omega_a \times \mathbf{r}_i) \cdot (\Omega_a \times \mathbf{r}_i)
\]

\[
= \sum_{i=1}^{N} m_i |\Omega_a \times \mathbf{r}_i|^2 = \sum_{i=1}^{N} m_i |\Omega_a|^2 |\mathbf{r}_i|^2 \sin^2 (\beta_i)
\]

where \( \beta_i \) denotes the angle between the position vector of the \( i^{th} \) point mass and the axis of the top. Since this expression is positive, this also shows \( \mathbf{L} \) has the same direction as \( \Omega_a \). Let \( \omega \equiv |\Omega_a| \). Then the above expression is of the form

\[
\mathbf{L} \cdot \Omega_a = I \omega^2,
\]

where

\[
I \equiv \sum_{i=1}^{N} m_i |\mathbf{r}_i|^2 \sin^2 (\beta_i).
\]

Thus, to get \( I \) you take the mass of the \( i^{th} \) point mass, multiply it by the square of its distance to the axis of the top and add all these up. This is defined as the moment of inertia of the top about the axis of the top. Letting \( \mathbf{u} \) denote a unit vector in the direction of the axis of the top, this implies

\[
\mathbf{L} = I \omega \mathbf{u}.
\]

(21.8)

Note the simple description of the angular momentum in terms of the moment of inertia. Referring to the above picture, define the vector, \( \mathbf{y} \) to be the projection of the vector, \( \mathbf{u} \) on the \( xy \) plane. Thus

\[
\mathbf{y} = \mathbf{u} - (\mathbf{u} \cdot \mathbf{k}) \mathbf{k}
\]

and

\[
(\mathbf{u} \cdot \mathbf{i}) = (\mathbf{y} \cdot \mathbf{i}) = \sin \alpha \cos \theta.
\]

(21.9)

Now also from (21.7),

\[
\frac{d\mathbf{L}}{dt} = \sum_{i=1}^{N} m_i \mathbf{r}_i' \times \mathbf{v}_i + m_i \mathbf{v}_i' = \sum_{i=1}^{N} \mathbf{r}_i \times m_i \mathbf{v}_i' = -\sum_{i=1}^{N} \mathbf{r}_i \times m_i g \mathbf{k}
\]
where $g$ is the acceleration of gravity. From (21.8), (21.9), and the above,

$$
\frac{d\mathbf{L}}{dt} \cdot \mathbf{i} = I\omega \left( \frac{du}{dt} \cdot \mathbf{i} \right) = I\omega \left( \frac{dy}{dt} \cdot \mathbf{i} \right)
= (-I\omega \sin \alpha \sin \theta) \theta' = -\sum_{i=1}^{N} \mathbf{r}_i \times m_i g \mathbf{k} \cdot \mathbf{i}
= -\sum_{i=1}^{N} m_i g \mathbf{r}_i \cdot \mathbf{k} \times \mathbf{i} = -\sum_{i=1}^{N} m_i g \mathbf{r}_i \cdot \mathbf{j}.
$$

(21.10)

To simplify this further, recall the following definition of the center of mass.

**Definition 21.4.1** Define the total mass, $M$ by

$$
M = \sum_{i=1}^{N} m_i
$$

and the center of mass, $\mathbf{r}_0$ by

$$
\mathbf{r}_0 \equiv \frac{\sum_{i=1}^{N} \mathbf{r}_i m_i}{M}.
$$

(21.11)

In terms of the center of mass, the last expression equals

$$
-Mg \mathbf{r}_0 \cdot \mathbf{j} = -Mg (\mathbf{r}_0 - (\mathbf{r}_0 \cdot \mathbf{k}) \mathbf{k} + (\mathbf{r}_0 \cdot \mathbf{k}) \mathbf{k} \cdot \mathbf{j} = -Mg (\mathbf{r}_0 - (\mathbf{r}_0 \cdot \mathbf{k}) \mathbf{k} \cdot \mathbf{j}
= -Mg |\mathbf{r}_0| - (\mathbf{r}_0 \cdot \mathbf{k}) \mathbf{k} \cos \theta
= -Mg |\mathbf{r}_0| \sin \alpha \cos \left( \frac{\pi}{2} - \theta \right).
$$

Note that by symmetry, $\mathbf{r}_0 (t)$ is on the axis of the top, is in the same direction as $\mathbf{L}$, $\mathbf{u}$, and $\mathbf{\Omega}_a$, and also $|\mathbf{r}_0|$ is independent of $t$. Therefore, from the second line of (21.10),

$$
(-I\omega \sin \alpha \sin \theta) \theta' = -Mg |\mathbf{r}_0| \sin \alpha \sin \theta.
$$

which shows

$$
\theta' = \frac{Mg |\mathbf{r}_0|}{I\omega}.
$$

(21.12)

From (21.12), the angular velocity of precession does not depend on $\alpha$ in the picture. It also is slower when $\omega$ is large and $I$ is large.

The above discussion is a considerable simplification of the problem of a spinning top obtained from an assumption that $\mathbf{\Omega}_a$ is approximately equal to $\mathbf{\Omega}$. It also leaves out all considerations of friction and the observation that the axis of symmetry wobbles. This is wobbling is called mutation. The full mathematical treatment of this problem involves the Euler angles and some fairly complicated differential equations obtained using techniques discussed in advanced physics classes. Lagrange studied these types of problems back in the 1700's.
21.4.2 Kinetic Energy

The next problem is that of understanding the total kinetic energy of a collection of moving point masses. Consider a possibly large number of point masses, \( m_i \) located at the positions \( r_i \) for \( i = 1, 2, \ldots, N \). Thus the velocity of the \( i^{th} \) point mass is \( \dot{r}_i = \mathbf{v}_i \). The kinetic energy of the mass \( m_i \) is defined by

\[
\frac{1}{2} m_i |\dot{r}_i|^2.
\]

(This is a very good time to review the presentation on kinetic energy given on Page 247.) The total kinetic energy of the collection of masses is then

\[
E = \sum_{i=1}^{N} \frac{1}{2} m_i |\dot{r}_i|^2. \tag{21.13}
\]

As these masses move about, so does the center of mass, \( \mathbf{r}_0 \). Thus \( \mathbf{r}_0 \) is a function of \( t \) just as the other \( r_i \). From (21.13) the total kinetic energy is

\[
E = \sum_{i=1}^{N} \frac{1}{2} m_i |\dot{r}_i - \dot{r}_0 + \dot{r}_0|^2
\]

\[
= \sum_{i=1}^{N} \frac{1}{2} m_i \left[ |\dot{r}_i - \dot{r}_0|^2 + |\dot{r}_0|^2 + 2 (\dot{r}_i - \dot{r}_0 \cdot \dot{r}_0) \right]. \tag{21.14}
\]

Now

\[
\sum_{i=1}^{N} m_i (\dot{r}_i - \dot{r}_0 \cdot \dot{r}_0) = \left( \sum_{i=1}^{N} m_i (r_i - \mathbf{r}_0) \right)' \cdot \dot{r}_0
\]

\[
= 0
\]

because from (21.11)

\[
\sum_{i=1}^{N} m_i (r_i - \mathbf{r}_0) = \sum_{i=1}^{N} m_i r_i - \sum_{i=1}^{N} m_i \mathbf{r}_0
\]

\[
= \sum_{i=1}^{N} m_i r_i - \sum_{i=1}^{N} m_i \left( \sum_{i=1}^{N} m_i r_i \right) = 0.
\]

Let \( M \equiv \sum_{i=1}^{N} m_i \) be the total mass. Then (21.14) reduces to

\[
E = \sum_{i=1}^{N} \frac{1}{2} m_i \left[ |\dot{r}_i|^2 + |\dot{r}_0|^2 \right]
\]

\[
= \frac{1}{2} M |\dot{r}_0|^2 + \sum_{i=1}^{N} \frac{1}{2} m_i |\dot{r}_i - \dot{r}_0|^2. \tag{21.15}
\]

The first term is just the kinetic energy of a point mass equal to the sum of all the masses involved, located at the center of mass of the system of masses while the second term represents kinetic energy which comes from the relative velocities of the masses taken with respect to the center of mass. It is this term which is considered more carefully in the case where the system of masses maintain distance between each other.
To illustrate the contrast between the case where the masses maintain a constant distance and on in which they don’t, take a hard boiled egg and spin it and then take a raw egg and give it a spin. You will certainly feel a big difference in the way the two eggs respond. Incidentally, this is a good way to tell whether the egg has been hard boiled or is raw and can be used to prevent messiness which could occur if you think it is hard boiled and it really isn’t.

Now let $e_1(t), e_2(t),$ and $e_3(t)$ be an orthonormal set of vectors which is fixed in the body undergoing rigid body motion. This means that $r_i(t) - r_0(t)$ has components which are constant in $t$ with respect to the vectors, $e_i(t)$. By Theorem 13.4.2 on Page 241 there exists a vector, $\Omega(t)$ which does not depend on $i$ such that

$$r'_i(t) - r'_0(t) = \Omega(t) \times (r_i(t) - r_0(t)).$$

Now using this in (21.15),

$$E = \frac{1}{2} M |r'_0|^2 + \frac{1}{2} \sum_{i=1}^{N} m_i |\Omega(t) \times (r_i(t) - r_0(t))|^2$$

$$= \frac{1}{2} M |r'_0|^2 + \frac{1}{2} \left( \sum_{i=1}^{N} m_i |r_i(t) - r_0(t)|^2 \sin^2 \theta_i \right) |\Omega(t)|^2$$

$$= \frac{1}{2} M |r'_0|^2 + \frac{1}{2} \left( \sum_{i=1}^{N} m_i |r_i(0) - r_0(0)|^2 \sin^2 \theta_i \right) |\Omega(t)|^2$$

where $\theta_i$ is the angle between $\Omega(t)$ and the vector, $r_i(t) - r_0(t)$. Therefore, $|r_i(t) - r_0(t)| \sin \theta_i$ is the distance between the point mass, $m_i$ located at $r_i$ and a line through the center of mass, $r_0$ with direction, $\Omega$ as indicated in the following picture.

Thus the expression, $\sum_{i=1}^{N} m_i |r_i(0) - r_0(0)|^2 \sin^2 \theta_i$ plays the role of a mass in the definition of kinetic energy except instead of the speed, substitute the angular speed, $|\Omega(t)|$. It is this expression which is called the moment of inertia about the line whose direction is $\Omega(t)$.

In both of these examples, the center of mass and the moment of inertia occurred in a natural way.

### 21.4.3 Finding The Moment Of Inertia And Center Of Mass

The methods used to evaluate multiple integrals make possible the determination of centers of mass and moments of inertia. In the case of a solid material rather than finitely many point masses, you replace the sums with integrals. The sums are essentially approximations of the integrals which result.

**Example 21.4.2** Let a solid occupy the three dimensional region $R$ and suppose the density is $\rho$. What is the moment of inertia of this solid about the $z$ axis? What is the center of mass?
Here the little masses would be of the form $\rho(x) \, dV$ where $x$ is a point of $R$. Therefore, the contribution of this mass to the moment of inertia would be

$$(x^2 + y^2) \rho(x) \, dV$$

where the Cartesian coordinates of the point $x$ are $(x, y, z)$. Then summing these up as an integral, yields the following for the moment of inertia.

$$\int_R (x^2 + y^2) \rho(x) \, dV.$$  \hfill (21.16)

To find the center of mass, sum up $r \rho \, dV$ for the points in $R$ and divide by the total mass. In Cartesian coordinates, where $r = (x, y, z)$, this means to sum up vectors of the form $(x \rho \, dV, y \rho \, dV, z \rho \, dV)$ and divide by the total mass. Thus the Cartesian coordinates of the center of mass are

$$\left( \frac{\int_R x \rho \, dV}{\int_R \rho \, dV}, \frac{\int_R y \rho \, dV}{\int_R \rho \, dV}, \frac{\int_R z \rho \, dV}{\int_R \rho \, dV} \right) \equiv \frac{\int_R r \rho \, dV}{\int_R \rho \, dV}.$$

Here is a specific example.

**Example 21.4.3.** Find the moment of inertia about the $z$ axis and center of mass of the solid which occupies the region, $R$ defined by $9 - (x^2 + y^2) \geq z \geq 0$ if the density is $\rho(x, y, z) = \sqrt{x^2 + y^2}$.

This moment of inertia is $\int_R (x^2 + y^2) \sqrt{x^2 + y^2} \, dV$ and the easiest way to find this integral is to use cylindrical coordinates. Thus the answer is

$$\int_0^{2\pi} \int_0^3 \int_0^{9-r^2} r^3 \, dz \, dr \, d\theta = \frac{8748}{35} \pi.$$

To find the center of mass, note the $x$ and $y$ coordinates of the center of mass,

$$\left( \frac{\int_R x \rho \, dV}{\int_R \rho \, dV}, \frac{\int_R y \rho \, dV}{\int_R \rho \, dV} \right)$$

both equal zero because the above shape is symmetric about the $z$ axis and $\rho$ is also symmetric in its values. Thus $x \rho \, dV$ will cancel with $-x \rho \, dV$ and a similar conclusion will hold for the $y$ coordinate. It only remains to find the $z$ coordinate of the center of mass, $\overline{z}$. In polar coordinates, $\rho = r$ and so,

$$\overline{z} = \frac{\int_R z \rho \, dV}{\int_R \rho \, dV} = \frac{\int_0^{2\pi} \int_0^3 \int_0^{9-r^2} rz^2 \, dz \, dr \, d\theta}{\int_0^{2\pi} \int_0^3 \int_0^{9-r^2} r^2 \, dz \, dr \, d\theta} = \frac{18}{7}.$$

Thus the center of mass will be $\left(0, 0, \frac{18}{7}\right)$.

### 21.5 Exercises

1. Let $R$ denote the finite region bounded by $z = 4 - x^2 - y^2$ and the $xy$ plane. Find $z_c$, the $z$ coordinate of the center of mass if the density, $\sigma$ is a constant.

2. Let $R$ denote the finite region bounded by $z = 4 - x^2 - y^2$ and the $xy$ plane. Find $z_c$, the $z$ coordinate of the center of mass if the density, $\sigma(x, y, z) = z$. 

3. Find the mass and center of mass of the region between the surfaces $z = -y^2 + 8$ and $z = 2x^2 + y^2$ if the density equals $\sigma = 1$.

4. Find the mass and center of mass of the region between the surfaces $z = -y^2 + 8$ and $z = 2x^2 + y^2$ if the density equals $\sigma (x, y, z) = x^2$.

5. The two cylinders, $x^2 + y^2 = 4$ and $y^2 + z^2 = 4$ intersect in a region, set, $R$. Find the mass and center of mass if the density, $\sigma$, is given by $\sigma (x, y, z) = z^2$.

6. The two cylinders, $x^2 + y^2 = 4$ and $y^2 + z^2 = 4$ intersect in a region, set, $R$. Find the mass and center of mass if the density equals $\sigma (x, y, z) = 4 + z$.

7. Find the mass and center of mass of the set, $(x, y, z)$ such that $x^2 + y^2 + z^2 \leq 1$ if the density is $\sigma (x, y, z) = 4 + y + z$.

8. Let $R$ denote the finite region bounded by $z = 9 - x^2 - y^2$ and the $xy$ plane. Find the moment of inertia of this shape about the $z$ axis.

9. Let $R$ denote the finite region bounded by $z = 9 - x^2 - y^2$ and the $xy$ plane. Find the moment of inertia of this shape about the $x$ axis.

10. Let $B$ be a solid ball of constant density and radius $R$. Find the moment of inertia about a line through a diameter of the ball. You should get $\frac{2}{5} R^2 M$.

11. Let $B$ be a solid ball of density, $\sigma = \rho$ where $\rho$ is the distance to the center of the ball which has radius $R$. Find the moment of inertia about a line through a diameter of the ball. Write your answer in terms of the total mass and the radius as was done in the constant density case.

12. Let $C$ be a solid cylinder of constant density and radius $R$. Find the moment of inertia about the axis of the cylinder

   You should get $\frac{1}{2} R^2 M$.

13. Let $C$ be a solid cylinder of constant density and radius $R$ and mass $M$ and let $B$ be a solid ball of radius $R$ and mass $M$. The cylinder and the sphere are placed on the top of an inclined plane and allowed to roll to the bottom. Which one will arrive first and why?

14. Suppose a solid of mass $M$ occupying the region, $B$ has moment of inertia, $I_l$ about a line, $l$ which passes through the center of mass of $M$ and let $l_1$ be another line parallel to $l$ and at a distance of $a$ from $l$. Then the parallel axis theorem states $I_{l_1} = I_l + a^2 M$. Prove the parallel axis theorem. Hint: Choose axes such that the $z$ axis is $l$ and $l_1$ passes through the point $(a, 0)$ in the $xy$ plane.

15. Using the parallel axis theorem find the moment of inertia of a solid ball of radius $R$ and mass $M$ about an axis located at a distance of $a$ from the center of the ball. Your answer should be $Ma^2 + \frac{2}{5} MR^2$.

16. Consider all axes in computing the moment of inertia of a solid. Will the smallest possible moment of inertia always result from using an axis which goes through the center of mass?

17. Find the moment of inertia of a solid thin rod of length $l$, mass $M$, and constant density about an axis through the center of the rod perpendicular to the axis of the rod. You should get $\frac{1}{12} l^2 M$. 

21.5. EXERCISES 401
18. Using the parallel axis theorem, find the moment of inertia of a solid thin rod of length \( l \), mass \( M \), and constant density about an axis through an end of the rod perpendicular to the axis of the rod. You should get \( \frac{1}{3} l^2 M \).

19. Let the angle between the \( z \) axis and the sides of a right circular cone be \( \alpha \). Also assume the height of this cone is \( h \). Find the \( z \) coordinate of the center of mass of this cone in terms of \( \alpha \) and \( h \) assuming the density is constant.

20. Let the angle between the \( z \) axis and the sides of a right circular cone be \( \alpha \). Also assume the height of this cone is \( h \). Assuming the density is \( \sigma = 1 \), find the moment of inertia about the \( z \) axis in terms of \( \alpha \) and \( h \).

21. Let \( R \) denote the part of the solid ball, \( x^2 + y^2 + z^2 \leq R^2 \) which lies in the first octant. That is \( x, y, z \geq 0 \). Find the coordinates of the center of mass if the density is constant. Your answer for one of the coordinates for the center of mass should be \( \frac{3}{8} R \).

22. Show that in general for \( \mathbf{L} \) angular momentum,

\[
\frac{d\mathbf{L}}{dt} = \mathbf{\Gamma}
\]

where \( \mathbf{\Gamma} \) is the total torque,

\[
\mathbf{\Gamma} = \sum \mathbf{r}_i \times \mathbf{F}_i
\]

where \( \mathbf{F}_i \) is the force on the \( i^{th} \) point mass.
The Integral On Other Sets

22.0.1 Outcomes
1. Define the $p$ dimensional volume.
2. Find the area of a surface.
3. Define and compute integrals over surfaces given parametrically.

22.1 The $p$ Dimensional Volume In $\mathbb{R}^n$

Consider the boundary of some three dimensional region such that a function, $f$ is defined on this boundary. Imagine taking the value of this function at a point, multiplying this value by the area of an infinitesimal chunk of area located at this point and then adding these up. This is just the notion of the integral presented earlier only now there is a difference because this infinitesimal chunk of area should be considered as two dimensional even though it is in three dimensions. However, it is not really all that different from what was done earlier. As before, it all depends on the following fundamental definition on Page 190.

Definition 22.1.1 Let $u_1, \ldots, u_p$ be vectors in $\mathbb{R}^n$. The $p$ dimensional parallelepiped determined by these vectors will be denoted by $P(u_1, \ldots, u_p)$ and it is defined as

$$P(u_1, \ldots, u_p) \equiv \left\{ \sum_{j=1}^{p} s_j u_j : s_j \in [0, 1] \right\}.$$ 

Define the volume of this parallelepiped by

$$\text{volume of } P(u_1, \ldots, u_p) \equiv (\det (u_i \cdot u_j))^{1/2}.$$ 

Suppose then that $x = f(u)$ where $u \in U$, a subset of $\mathbb{R}^p$ and $x$ is a point in $V$, a subset of $n$ dimensional space where $n \geq p$. Thus, letting the Cartesian coordinates of $x$ be given by $x = (x_1, \ldots, x_n)^T$, each $x_i$ being a function of $u$, an infinitesimal box located at $u_0$ corresponds to an infinitesimal parallelepiped located at $f(u_0)$ which is determined by the $p$ vectors $\left\{ \frac{\partial x(u_0)}{\partial u_i} \, du_i \right\}_{i=1}^p$, each of which is tangent to the surface defined by $x = f(u)$. (No sum on the repeated index.) From Definition 22.1.1, the volume of this infinitesimal parallelepiped located at $f(u_0)$ is given by

$$\det \left( \frac{\partial x(u_0)}{\partial u_i} \, du_i, \frac{\partial x(u_0)}{\partial u_j} \, du_j \right)^{1/2}.$$ 

(22.1)
I like to think of this in the case where $p = 2$. In this case the infinitesimal parallelepiped is an infinitesimal parallelogram tangent to the surface defined by $x = f(u)$ like a very small scale on a lizard. This is the essence of the idea. To define the area of the lizard sum up areas of individual scales.

Now, continuing with the general case, the matrix in the above formula is a $p \times p$ matrix. Denoting

$$\frac{\partial x(u_0)}{\partial u_i} = x_{i,j}$$

to save space, this matrix is of the form

$$\begin{pmatrix}
\begin{bmatrix}
du_1 & 0 & \cdots & 0 \\
0 & du_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & du_p
\end{bmatrix}
&
\begin{bmatrix}
x_{1,1} & \cdots & \cdots & x_{1,p} \\
x_{2,1} & \cdots & \cdots & x_{2,p} \\
\vdots & \vdots & \ddots & \vdots \\
x_{p,1} & \cdots & \cdots & x_{p,p}
\end{bmatrix}

\end{pmatrix}$$

$$\begin{pmatrix}
\begin{bmatrix}
x_{1} & x_{2} & \cdots & x_{p}
\end{bmatrix}
&
\begin{bmatrix}
du_1 & 0 & \cdots & 0 \\
0 & du_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & du_p
\end{bmatrix}

\end{pmatrix}$$

Therefore, by the theorem which says the determinant of a product equals the product of the determinants, the determinant of the above product equals

$$\det \left( \begin{bmatrix}
x_{1} & x_{2} & \cdots & x_{p}
\end{bmatrix}
\right) \left( \begin{bmatrix}
\begin{bmatrix}
x_{1} & x_{2} & \cdots & x_{p}
\end{bmatrix}
&
\begin{bmatrix}
du_1 & 0 & \cdots & 0 \\
0 & du_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & du_p
\end{bmatrix}

\end{pmatrix} \right)^2 \left( du_1 du_2 \cdots du_p \right)^2$$

and so taking the square root implies the volume of the infinitesimal parallelepiped at $x = f(u_0)$ is

$$\det \left( \frac{\partial x(u_0)}{\partial u_i} \cdot \frac{\partial x(u_0)}{\partial u_j} \right)^{1/2} du_1 du_2 \cdots du_p = \det \left( Df(u)^T Df(u) \right)^{1/2} du_1 du_2 \cdots du_p$$

**Definition 22.1.2** Let $x = f(u)$ be as described above. Then the symbol, $\frac{\partial(x_1, \cdots, x_n)}{\partial(u_1, \cdots, u_p)}$, is defined by

$$\det \left( \frac{\partial x(u_0)}{\partial u_i} \cdot \frac{\partial x(u_0)}{\partial u_j} \right)^{1/2} \equiv \frac{\partial(x_1, \cdots, x_n)}{\partial(u_1, \cdots, u_p)}.$$  

Also, the symbol, $dV_p \equiv \frac{\partial(x_1, \cdots, x_n)}{\partial(u_1, \cdots, u_p)} du_1 \cdots du_p$ is called the volume element or area element. Note the use of the subscript, $p$. This indicates the $p$ dimensional volume element. When $p = 2$ it is customary to write $dA$. Also, continue referring to $\frac{\partial(x_1, \cdots, x_n)}{\partial(u_1, \cdots, u_p)}$ as the Jacobian.
This motivates the following fundamental procedure which I hope is extremely familiar from the earlier material.

**Procedure 22.1.3** Suppose $U$ is a subset of $\mathbb{R}^p$ and suppose $f : U \rightarrow f(U) \subseteq \mathbb{R}^n$ is a one to one and $C^1$ function. Then if $h : f(U) \rightarrow \mathbb{R}$, define the $p$ dimensional surface integral, $\int_{f(U)} h(x) \, dV_p$ according to the following formula.

$$\int_{f(U)} h(x) \, dV_p \equiv \int_U h(f(u)) \frac{\partial (x_1, \ldots, x_n)}{\partial (u_1, \ldots, u_p)} \, dV.$$ 

**Example 22.1.4** Find the area of the region labeled $A$ in the following picture. The two circles are of radius 1, one has center $(0, 0)$ and the other has center $(1, 0)$.

![Diagram showing two circles](image)

The circles bounding these disks are $x^2 + y^2 = 1$ and $(x - 1)^2 + y^2 = x^2 + y^2 - 2x + 1 = 1$. Therefore, in polar coordinates these are of the form $r = 1$ and $r = 2 \cos \theta$.

The set $A$ corresponds to the set $U$, in the $(\theta, r)$ plane determined by $\theta \in [-\pi/3, \pi/3]$ and for each value of $\theta$ in this interval, $r$ goes from 1 up to $2 \cos \theta$. Therefore, the area of this region is of the form,

$$\int_U 1 \, dV = \int_{-\pi/3}^{\pi/3} \int_1^{2 \cos \theta} \frac{\partial (x_1, x_2)}{\partial (\theta, r)} \, dr \, d\theta.$$ 

It is necessary to find $\frac{\partial (x_1, x_2)}{\partial (\theta, r)}$. The mapping $f : U \rightarrow \mathbb{R}^2$ takes the form $f(\theta, r) = (r \cos \theta, r \sin \theta)^T$ and so

$$Df(\theta, r) = \begin{pmatrix} -r \sin \theta & \cos \theta \\ r \cos \theta & \sin \theta \end{pmatrix},$$

and so

$$Df(\theta, r)^T Df(\theta, r) = \begin{pmatrix} r^2 & 0 \\ 0 & 1 \end{pmatrix},$$

which implies

$$\frac{\partial (x_1, x_2)}{\partial (\theta, r)} = \det \left( Df(\theta, r)^T Df(\theta, r) \right)^{1/2} = r.$$ 

Therefore, the area element is $r \, dr \, d\theta$. It follows the desired area is

$$\int_{-\pi/3}^{\pi/3} \int_1^{2 \cos \theta} r \, dr \, d\theta = \frac{1}{2} \sqrt{3} + \frac{1}{3} \pi.$$ 

**Example 22.1.5** Consider the surface given by $z = x^2$ for $(x, y) \in [0, 1] \times [0, 1] = U$. Find the surface area of this surface.
The first step in using the above is to write this surface in the form \( x = f(u) \). This is easy to do if you let \( u = (x, y) \). Then \( f(x, y) = (x, y, x^2) \). If you like, let \( x = u_1 \) and \( y = u_2 \).

What is \( \frac{\partial (x_1, x_2, x_3)}{\partial (x,y)} \)?

\[
Df(x, y) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2x & 0 \end{pmatrix}
\]

and so

\[
Df(x, y)^T Df(u) = \begin{pmatrix} 1 + 4x^2 & 0 \\ 0 & 1 \end{pmatrix}
\]

Thus in this case,

\[
Df(u)^T Df(u) = \begin{pmatrix} 1 + 4x^2 & 0 \\ 0 & 1 \end{pmatrix}
\]

and so the area element is \( \sqrt{1 + 4x^2} \, dx \, dy \) and the surface area is obtained by integrating the function, \( h(x) \equiv 1 \). Therefore, this area is

\[
\int_U dV = \int_0^1 \int_0^1 \sqrt{1 + 4x^2} \, dx \, dy = \frac{1}{2} \sqrt{5} - \frac{1}{4} \ln \left( -2 + \sqrt{5} \right)
\]

which can be obtained by using the trig. substitution, \( 2x = \tan \theta \) on the inside integral.

Note this all depends on being able to write the surface in the form, \( x = f(u) \) for \( u \in U \subseteq \mathbb{R}^p \). Surfaces obtained in this form are called parametrically defined surfaces. These are best but sometimes you have some other description of a surface and in these cases things can get pretty intractable. For example, you might have a level surface of the form \( 3x^2 + 4y^4 + z^6 = 10 \). In this case, you could solve for \( z \) using methods of algebra. Thus \( z = \sqrt[6]{10 - 3x^2 - 4y^4} \) and a parametric description of part of this level surface is \( (x, y, \sqrt[6]{10 - 3x^2 - 4y^4}) \) for \( (x, y) \in U \) where \( U = \{(x, y) : 3x^2 + 4y^4 \leq 10\} \). But what if the level surface was something like

\[
\sin \left( x^2 + \ln \left( 7 + y^2 \sin x \right) \right) + \sin (zx) e^z = 11 \sin (xyz)?
\]

I really don’t see how to use methods of algebra to solve for some variable in terms of the others. It isn’t even clear to me whether there are any points \( (x, y, z) \in \mathbb{R}^3 \) satisfying this particular relation. However, if a point satisfying this relation can be identified, the implicit function theorem from advanced calculus can usually be used to assert one of the variables is a function of the others, proving the existence of a parameterization at least locally. However, this theorem doesn’t give us the answer in terms of known functions so this isn’t much help. Finding a parametric description of a surface is a hard problem and there are no easy answers.

**Example 22.1.6** Let \( U = [0, 12] \times [0, 2\pi] \) and let \( f : U \to \mathbb{R}^3 \) be given by \( f(t, s) \equiv (2\cos t + \cos s, 2\sin t + \sin s, t)^T \). Find a double integral for the surface area. A graph of this surface is drawn below.
It looks like something you would use to make sausages. Anyway,

\[ Df(t, s) = \begin{pmatrix} -2\sin t & -\sin s \\ 2\cos t & \cos s \\ 1 & 0 \end{pmatrix} \]

and so

\[ Df(t, s)^T Df(t, s) = \begin{pmatrix} 5 & 2\sin t\sin s + 2\cos t\cos s \\ 2\sin t\sin s + 2\cos t\cos s & 1 \end{pmatrix} \]

which implies the area equals

\[
\int_0^{2\pi} \int_0^{2\pi} \sqrt{5 - 4\sin^2 t\sin^2 s - 8\sin t\sin s\cos t\cos s - 4\cos^2 t\cos^2 s} \, dt \, ds.
\]

If you really needed to find the number this equals, how would you go about finding it? This is an interesting question and there is no single right answer. You should think about this. Here is an example for which you will be able to find the integrals.

**Example 22.1.7** Let \( U = [0, 2\pi] \times [0, 2\pi] \) and for \( (t, s) \in U \), let

\[ f(t, s) = (2\cos t + \cos t\cos s, -2\sin t - \sin t\cos s, \sin s)^T. \]

Find the area of \( f(U) \). This is the surface of a donut shown below. The fancy name for this shape is a torus.
To find its area,
\[
Df(t, s) = \begin{pmatrix}
-2\sin t - \sin t \cos s & -\cos t \sin s \\
-2\cos t - \cos t \cos s & \sin t \sin s \\
0 & \cos s
\end{pmatrix}
\]
and so
\[
Df(t, s)^T Df(t, s) = \begin{pmatrix}
4 + 4\cos s + \cos^2 s & 0 \\
0 & 1
\end{pmatrix}
\]
which implies the area element is
\[
\det \left( \begin{array}{cc}
4 + 4\cos s + \cos^2 s & 0 \\
0 & 1
\end{array} \right) \frac{1}{2} \, ds \, dt = (4 + 4\cos s + \cos^2 s)^{1/2} \, ds \, dt
\]
and the area is
\[
\int_0^{2\pi} \int_0^{2\pi} (\cos s + 2) \, ds \, dt = 8\pi^2
\]

**Example 22.1.8** Let \( U = [0, 2\pi] \times [0, 2\pi] \) and for \((t, s) \in U\), let
\[
f(t, s) = (2\cos t + \cos t \cos s, -2\sin t - \sin t \cos s, \sin s)^T.
\]
Find
\[
\int_{f(U)} h \, dV
\]
where \( h(x, y, z) = x^2 \).

Everything is the same as the preceding example except this time it is an integral of a function. The area element is \((\cos s + 2) \, ds \, dt\) and so the integral called for is
\[
\int_{f(U)} h \, dV = \int_0^{2\pi} \int_0^{2\pi} x^2 \text{ on the surface} \frac{x^2}{2\cos t + \cos t \cos s} (\cos s + 2) \, ds \, dt = 22\pi^2
\]

**Example 22.1.9** Let \( U = \{(x, y, z) : x^2 + y^2 + z^2 \leq 4\} \) and for \((x, y, z) \in U\) let \( f(x, y, z) = (x, y, x + y, z) \). Find the three dimensional volume of \( f(U) \).
Note there is no picture here because I am unable to draw one in four dimensions. Nevertheless it is a three dimensional volume which is being computed. Everything is done the same as before.

\[
Df(x, y, z) = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
1 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]

and so

\[
Df(x, y, z)^T Df(x, y, z) = \begin{pmatrix}
2 & 1 & 0 \\
1 & 2 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]

and so the volume element is \(3 \, dx \, dy \, dz\). Therefore, the volume of \(f(U)\) is

\[
\int_U 3 \, dx \, dy \, dz = 3 \left( \frac{4}{3} \pi (8) \right) = 32 \pi.
\]

The special case where a surface is in the form \(z = f(x, y), (x, y) \in U\), yields a simple formula which is used most often in this situation. You write the surface parametrically in the form \(f(x, y) = (x, y, f(x, y))^T : (x, y) \in U\). Then

\[
Df(x, y) = \begin{pmatrix}
1 & 0 \\
0 & 1 \\
f_x & f_y
\end{pmatrix}
\]

and so

\[
Df(x, y, z)^T Df(x, y, z) = \begin{pmatrix}
1 + f_x^2 & f_x f_y \\
f_x f_y & 1 + f_y^2
\end{pmatrix}.
\]

Thus,

\[
det \left( \begin{pmatrix}
1 + f_x^2 & f_x f_y \\
f_x f_y & 1 + f_y^2
\end{pmatrix} \right) = 1 + f_x^2 + f_y^2
\]

and so the area element is \(\sqrt{1 + f_x^2 + f_y^2} \, dx \, dy\).

When the surface of interest comes in this simple form, people generally use this area element directly rather than worrying about a parameterization and taking determinants and finding matrices.

In the case where the surface is of the form \(x = f(y, z)\) for \((y, z) \in U\), the area element is obtained similarly and is

\[
\sqrt{1 + f_y^2 + f_z^2} \, dy \, dz.
\]

I think you can guess what the area element is if \(y = f(x, z)\).

There is also a simple geometric description of these area elements. Consider the surface \(z = f(x, y)\). This is a level surface of the function of three variables \(z - f(x, y)\). In fact the surface is simply \(z - f(x, y) = 0\). Now consider the gradient of this function of three variables. The gradient is perpendicular to the surface and the third component is positive in this case. This gradient is \((- f_x, - f_y, 1)\) and so the unit upward normal is just \(\frac{1}{\sqrt{f_x^2 + f_y^2}} (- f_x, - f_y, 1)\).

Now consider the following picture.
In this picture, you are looking at a chunk of area on the surface seen on edge and so it seems reasonable to expect to have \( d\theta \, dx \, dy = dV \cos \theta \). But it is easy to find \( \cos \theta \) from the picture and the properties of the dot product.

\[
\cos \theta = \frac{n \cdot k}{|n||k|} = \frac{1}{\sqrt{1 + f_x^2 + f_y^2}}
\]

Therefore, \( dV = \sqrt{1 + f_x^2 + f_y^2} \, dx \, dy \) as claimed. In this context, the surface involved is referred to as \( S \) because the vector valued function, \( f \) giving the parameterization will not have been identified.

**Example 22.1.10** Let \( z = \sqrt{x^2 + y^2} \) where \( (x, y) \in U \) for \( U = \{(x, y) : x^2 + y^2 \leq 4\} \) Find

\[
\int_S h \, dV
\]

where \( h(x, y, z) = x + z \) and \( S \) is the surface described as \( (x, y, \sqrt{x^2 + y^2}) \) for \( (x, y) \in U \).

Here you can see directly the angle in the above picture is \( \frac{\pi}{4} \) and so \( dV = \sqrt{2} \, dx \, dy \). If you don’t see this or if it is unclear, simply compute \( \sqrt{1 + f_x^2 + f_y^2} \) and you will find it is \( \sqrt{2} \). Therefore, using polar coordinates,

\[
\int_S h \, dV = \int_U \left( x + \sqrt{x^2 + y^2} \right) \sqrt{2} \, dV
\]

\[
= \sqrt{2} \int_0^{2\pi} \int_0^\infty (r \cos \theta + r) \, r \, dr \, d\theta
\]

\[
= \frac{16}{3} \sqrt{2}\pi.
\]

One other issue is worth mentioning. Suppose \( f_i : U_i \rightarrow \mathbb{R}^p \) where \( U_i \) are sets in \( \mathbb{R}^p \) and suppose \( f_1 (U_1) \) intersects \( f_2 (U_2) \) along \( C \) where \( C = h(V) \) for \( V \subseteq \mathbb{R}^k \) for \( k < p \). Then define integrals and areas over \( f_1 (U_1) \cup f_2 (U_2) \) as follows.

\[
\int_{f_1(U_1) \cup f_2(U_2)} g \, dV_p \equiv \int_{f_1(U_1)} g \, dV_p + \int_{f_2(U_2)} g \, dV_p.
\]

Admittedly, the set \( C \) gets added in twice but this doesn’t matter because its \( p \) dimensional volume equals zero and therefore, the integrals over this set will also be zero. Why is this? To find the \( p \) dimensional volume element on \( C \), it is necessary to find a function, \( f \), mapping \( U \subseteq \mathbb{R}^p \) to \( C \). Let \( f(v, s_1, \cdots, s_{p-k}) \equiv h(v) \). Then \( Df(v, s_1, \cdots, s_{p-k}) \) has at least one column of zeros and so \( \det \left( Df^T \, Df \right) = 0 \) showing the \( p \) dimensional volume element is zero and so this makes no contribution to the integral as claimed. Clearly something similar holds in the case of many surfaces joined in this way.
I have been purposely vague about precise mathematical conditions necessary for the above procedures. This is because the precise mathematical conditions which are usually cited are very technical and at the same time far too restrictive. The most general conditions under which these sorts of procedures are valid include things like Lipschitz functions defined on very general sets. These are functions satisfying a Lipschitz condition of the form \(|f(x) - f(y)| \leq K|x - y|\). For example, \(y = |x|\) is Lipschitz continuous. However, this function does not have a derivative at every point. So it is with Lipschitz functions. However, it turns out these functions have derivatives at enough points to push everything through but this requires considerations involving the Lebesgue integral. Lipschitz functions are also not the most general kind of function for which the above is valid.

22.2 Spherical Coordinates In \(\mathbb{R}^n\)

Recall polar coordinates are of the form
\[
\begin{align*}
  x_1 &= \rho \cos \theta \\
  x_2 &= \rho \sin \theta
\end{align*}
\]
where \(\rho > 0\) and \(\theta \in [0, 2\pi)\). Here I am writing \(\rho\) in place of \(r\) to emphasize a pattern which is about to emerge. I will consider polar coordinates as spherical coordinates in two dimensions. This is also the reason I am writing \(x_1\) and \(x_2\) instead of the more usual \(x\) and \(y\). Now consider what happens when you go to three dimensions. The situation is depicted in the following picture.

From this picture, you see that \(x_3 = \rho \cos \phi_1\). Also the distance between \((x_1, x_2)\) and \((0, 0)\) is \(\rho \sin (\phi_1)\). Therefore, using polar coordinates to write \((x_1, x_2)\) in terms of \(\theta\) and this distance,
\[
\begin{align*}
  x_1 &= \rho \sin \phi_1 \cos \theta, \\
  x_2 &= \rho \sin \phi_1 \sin \theta, \\
  x_3 &= \rho \cos \phi_1
\end{align*}
\]
where \(\phi_1 \in [0, \pi]\). What was done is to replace \(\rho\) with \(\rho \sin \phi_1\) and then to add in \(x_3 = \rho \cos \phi_1\). Having done this, there is no reason to stop with three dimensions. Consider the following picture:

From this picture, you see that \(x_4 = \rho \cos \phi_2\). Also the distance between \((x_1, x_2, x_3)\) and \((0, 0, 0)\) is \(\rho \sin (\phi_2)\). Therefore, using polar coordinates to write \((x_1, x_2, x_3)\) in terms of \(\theta, \phi_1, \phi_2\),
and this distance,
\[
x_1 = \rho \sin \phi_2 \sin \phi_1 \cos \theta,
\]
\[
x_2 = \rho \sin \phi_2 \sin \phi_1 \sin \theta,
\]
\[
x_3 = \rho \sin \phi_2 \cos \phi_1,
\]
\[
x_4 = \rho \cos \phi_2,
\]
where \(\phi_2 \in [0, \pi]\).

Continuing this way, given spherical coordinates in \(\mathbb{R}^n\), to get the spherical coordinates in \(\mathbb{R}^{n+1}\), you let \(x_{n+1} = \rho \cos \phi_{n-1}\) and then replace every occurrence of \(\rho\) with \(\rho \sin \phi_{n-1}\) to obtain \(x_1 \cdots x_n\) in terms of \(\phi_1, \phi_2, \ldots, \phi_{n-1}, \theta, \) and \(\rho\).

For spherical coordinates, it is always the case that \(\rho\) measures the distance from the point in \(\mathbb{R}^n\) to the origin in \(\mathbb{R}^n\), \(0\). Each \(\phi_i \in [0, \pi]\), and \(\theta \in [0, 2\pi]\). I leave it as an exercise using math induction to prove that these coordinates map \(\prod_{i=1}^{n-2} [0, \pi] \times [0, 2\pi] \times (0, \infty)\) one to one onto \(\mathbb{R}^n \setminus \{0\}\).

It is customary to write \(S^{n-1}\) for the set \(\{x \in \mathbb{R}^n : |x| = 1\}\). Thus a parameterization for this level surface is given by letting \(\rho = 1\) in spherical coordinates. I will denote by \(S^{n-1}(a)\) the sphere having radius \(a > 0\). What would the \(n-1\) dimensional volume element on \(S^{n-1}(a)\) be? For \(S^1(a)\), there is only one parameter, \(\theta\). Therefore, the one dimensional volume element is
\[
det \left( \begin{pmatrix} \sin \theta & a \cos \theta \\ a \cos \theta & a \cos \theta \end{pmatrix} \right) = a d\theta.
\]
where \(\theta \in [0, 2\pi]\).

Next consider \(S^2\). In this case the two dimensional volume element is
\[
det \left( \begin{pmatrix} -\sin \phi_1 & \sin \phi_1 & 0 \\ \cos \phi_1 & \cos \phi_1 & \sin \phi_1 \\ 0 & -\sin \phi_1 & \sin \phi_1 \end{pmatrix} \right) \frac{1}{2} d\phi_1 d\theta = a^2 (\sin \phi_1) d\phi_1 d\theta.
\]

What of \(S^n(a)\) \(\equiv \{x \in \mathbb{R}^n : |x| = a\}\) and \(S^0(a) = \{x \in \mathbb{R}^{n+1} : |x| = a\}\). Let \(x_n\) denote the vector \((x_1, \ldots, x_n)\). That is, \(x_n\) consists of the first \(n\) components of \(x\). Let \(D_{\theta_{\phi_{n-2}}}\) denote the derivative with respect to the vector, \((\theta, \phi_1 \cdots \phi_{n-2})\). Then the volume element on \(S^n(a)\) is of the form
\[
det \left( \begin{pmatrix} \sin (\phi_{n-1}) (D_{\theta_{\phi_{n-2}}} x_n)_{(n-1)\times n}^T & 0 \\ 0 & -a \sin \phi_{n-1} \end{pmatrix} \right) \frac{1}{2} d\phi_{n-1} \cdots d\phi_1 d\theta
\]
Now using block multiplication, this reduces to
\[
a \sin (\phi_{n-1}) \det \left( \begin{pmatrix} (D_{\theta_{\phi_{n-2}}} x_n)_{(n-1)\times n}^T & (D_{\theta_{\phi_{n-2}}} x_n)_{(n-1)\times n} \end{pmatrix} \right) \frac{1}{2} d\phi_{n-1} \cdots d\phi_1 d\theta.
\]
That is, to get the volume element in $S^n(a)$, you multiply the volume element on $S^{n-1}(a)$ by $a \sin(\phi_{n-1}) \, d\phi_{n-1}$. Consequently, beginning with the volume element on $S^1(a)$, you obtain the succession of volume elements for $S^1(a), S^2(a), S^3(a), S^4(a)$.

$$\int a \sin(\phi_{n-1}) \, d\phi_{n-1}, a^2 \sin(\phi_1) \, d\phi_1 \, d\theta, a^3 \sin(\phi_2) \, d\phi_2 \, d\phi_1 \, d\theta,$$

$$a^4 \sin(\phi_3) \, d\phi_3 \, d\phi_2 \, d\phi_1 \, d\theta, \text{ etc.}$$

In general, the $n$ dimensional volume element on $S^n(a)$ is

$$a^n \left( \prod_{i=1}^{n-1} \sin(\phi_i) \right) \left( \prod_{i=1}^{n-1} d\phi_i \right) \, d\theta \quad (22.2)$$

Using similar reasoning, the $n$ dimensional volume element in terms of the spherical coordinates is

$$\rho^{n-1} \left( \prod_{i=1}^{n-2} \sin(\phi_i) \right) \left( \prod_{i=1}^{n-2} d\phi_i \right) \, d\rho \, d\phi,$$  

$$\rho^{n-1} \left( \prod_{i=1}^{n-2} \sin(\phi_i) \right) \left( \prod_{i=1}^{n-2} d\phi_i \right) \, d\rho \, d\phi \quad (22.3)$$

Formulas (22.2) and (22.3) are very useful in estimating integrals.

**Example 22.2.1** For what values of $s$ is the integral $\int_{B(0,R)} (1 + |x|^2)^s \, dV$ bounded independent of $R$? Here $B(0,R)$ is the ball, $\{x \in \mathbb{R}^n : |x| \leq R\}$.

I think you can see immediately that $s$ must be negative but exactly how negative? It turns out it depends on $n$ and using spherical coordinates, you can find just exactly what is needed. It is very hard to overstate the importance of the technique I am about to show you. Write the above integral in $n$ dimensional spherical coordinates.

$$\int_0^R \int_0^\pi \cdots \int_0^\pi \int_0^{2\pi} \rho^{n-1} (1 + \rho^2)^s \left( \prod_{i=1}^{n-1} \sin(\phi_i) \right) \left( \prod_{i=1}^{n-1} d\phi_i \right) \, d\theta \, d\rho$$

$$= \int_0^R \rho^{n-1} (1 + \rho^2)^s \int_{S^{n-1}} dS^{n-1} \, d\rho = \omega_n \int_0^R \rho^{n-1} (1 + \rho^2)^s \, d\rho$$

where $dS^{n-1} = \left( \prod_{i=1}^{n-2} \sin(\phi_i) \right) \left( \prod_{i=1}^{n-2} d\phi_i \right) \, d\theta$.

$$\omega_n \equiv \int_0^\pi \cdots \int_0^\pi \int_0^{2\pi} \left( \prod_{i=1}^{n-2} \sin(\phi_i) \right) \left( \prod_{i=1}^{n-2} d\phi_i \right) \, d\theta$$

and from the above explanation this equals the area of $S^{n-1}$, the unit sphere, $\{x \in \mathbb{R}^n : |x| = 1\}$. Now the very hard problem has been reduced to considering an easy one variable problem of finding when

$$\int_0^R \rho^{n-1} (1 + \rho^2)^s \, d\rho$$

is bounded independent of $R$. I leave it to you to verify using standard one variable calculus that you need $2s + (n-1) < -1$ so you need $s < -n/2$.

### 22.3 Exercises With Answers

1. Find a parameterization for the intersection of the planes $x + y + 2z = -3$ and $2x - y + z = -4$.

   **Answer:**
   $$(x, y, z) = \left( -t - \frac{7}{4}, -t - \frac{2}{4}, t \right)$$
2. Find a parameterization for the intersection of the plane $4x + 2y + 4z = 0$ and the circular cylinder $x^2 + y^2 = 16$.

Answer:

The cylinder is of the form $x = 4\cos t, y = 4\sin t$ and $z = z$. Therefore, from the equation of the plane, $16\cos t + 8\sin t + 4z = 0$. Therefore, $z = -16\cos t - 8\sin t$ and this shows the parameterization is of the form $(x, y, z) = (4\cos t, 4\sin t, -16\cos t - 8\sin t)$ where $t \in [0, 2\pi]$.

3. Find a parameterization for the intersection of the plane $3x + 2y + z = 4$ and the elliptic cylinder $x^2 + 4z^2 = 1$.

Answer:

The cylinder is of the form $x = \cos t, 2z = \sin t$ and $y = y$. Therefore, from the equation of the plane, $3\cos t + 2y + \frac{1}{2}\sin t = 4$. Therefore, $y = 2 - \frac{3}{2}\cos t - \frac{1}{4}\sin t$ and this shows the parameterization is of the form $(x, y, z) = (\cos t, 2 - \frac{3}{2}\cos t - \frac{1}{4}\sin t, \frac{1}{2}\sin t)$ where $t \in [0, 2\pi]$.

4. Find a parameterization for the straight line joining $(4, 3, 2)$ and $(1, 7, 6)$.

Answer:

$$(x, y, z) = (4, 3, 2) + t(-3, 4, 4) = (4 - 3t, 3 + 4t, 2 + 4t)$$ where $t \in [0, 1]$.

5. Find a parameterization for the intersection of the surfaces $y + 3z = 4x^2 + 4$ and $4y + 4z = 2x + 4$.

Answer:

This is an application of Cramer’s rule. $y = -2x^2 - \frac{1}{3} + \frac{3}{4}x, z = -\frac{1}{3}x + \frac{3}{2} + 2x^2$. Therefore, the parameterization is $(x, y, z) = (t, -2t^2 - \frac{1}{3} + \frac{3}{4}t, -\frac{1}{3}t + \frac{3}{2} + 2t^2)$.

6. Find the area of $S$ if $S$ is the part of the circular cylinder $x^2 + y^2 = 16$ which lies between $z = 0$ and $z = 4 + y$.

Answer:

Use the parameterization, $x = 4\cos v, y = 4\sin v$ and $z = u$ with the parameter domain described as follows. The parameter, $v$ goes from $-\frac{\pi}{2}$ to $\frac{3\pi}{2}$ and for each $v$ in this interval, $u$ should go from $0$ to $4 + 4\sin v$. To see this observe that the cylinder has its axis parallel to the $z$ axis and if you look at a side view of the surface you would see something like this:
The positive x axis is coming out of the paper toward you in the above picture and the angle \( v \) is the usual angle measured from the positive x axis. Therefore, the area is just 
\[
A = \int_{\pi/2}^{3\pi/2} \int_0^4 4 + 4 \sin v \, du \, dv = 32\pi.
\]

7. Find the area of \( S \) if \( S \) is the part of the cone \( x^2 + y^2 = 9z^2 \) between \( z = 0 \) and \( z = h \).

Answer:

When \( z = h \), \( x^2 + y^2 = 9h^2 \) which is the boundary of a circle of radius \( ah \). A parameterization of this surface is 
\[
x = u, \quad y = v, \quad z = \frac{1}{3} \sqrt{(9h^2 - u^2)}
\]
where \((u, v) \in D\), a disk centered at the origin having radius \( ha \). Therefore, the volume is just 
\[
\int_D \sqrt{1 + z_u^2 + z_v^2} \, dA = \int_{-ha}^{ha} \int_{\sqrt{(9h^2 - u^2)}^{1/3}} \sqrt{10} \, dv \, du = 3\pi h^2 \sqrt{10}
\]

8. Parametrizing the cylinder \( x^2 + y^2 = 4 \) by \( x = 2 \cos v, y = 2 \sin v, z = u \), show that the area element is \( dA = 2 \, du \, dv \).

Answer:

It is necessary to compute \( \frac{\partial(x, y, z)}{\partial(u, v)} = \det(DF^T DF) \).

\[
DF(u, v) = \begin{pmatrix} 0 & -2 \sin v \\ 0 & 2 \cos v \\ 1 & 0 \end{pmatrix}
\]

and so \( DF^T DF = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} \) and so the area element is as described.

9. Find the area enclosed by the limacon \( r = 2 + \cos \theta \).

Answer:

You can graph this region and you see it is sort of an oval shape and that \( \theta \in [0, 2\pi] \) while \( r \) goes from 0 up to \( 2 + \cos \theta \). Now \( x = r \cos \theta \) and \( y = r \sin \theta \) are the \( x \) and \( y \) coordinates corresponding to \( r \) and \( \theta \) in the above parameter domain. Therefore, the area of the limacon equals 
\[
\int \int_D \left| \frac{\partial(x, y)}{\partial(\theta, \phi)} \right| \, dr \, d\theta = \int_0^{2\pi} \int_0^{2 + \cos \theta} r \, dr \, d\theta
\]
because the Jacobian equals \( r \) in this case. Therefore, the area equals 
\[
\int_0^{2\pi} \int_0^{2 + \cos \theta} r \, dr \, d\theta = \frac{9}{2}\pi.
\]

10. Find the surface area of the paraboloid \( z = h(1 - x^2 - y^2) \) between \( z = 0 \) and \( z = h \).

Answer:
Let \( R \) denote the unit circle. Then the area of the surface above this circle would be
\[
\int \int_R \sqrt{1 + 4x^2 + 4y^2} \, dA.
\]
Changing to polar coordinates, this becomes
\[
\int_0^2 \pi \int_0^1 \sqrt{1 + 4r^2} \, dr \, d\theta = \frac{1}{6} \pi \sqrt{(1+4h^2) + 4(1+4h^2)h^2-1}.
\]

11. Evaluate \( \int \int_S (1 + x) \, dA \) where \( S \) is the part of the plane \( 2x + 3y + 3z = 18 \) which is in the first octant.

Answer:
\[
\int_0^6 \int_0^{6 - \frac{x}{2}} (1 + x) \frac{1}{3} \sqrt{22} \, dy \, dx = 28 \sqrt{22}
\]

12. Evaluate \( \int \int_S (1 + x) \, dA \) where \( S \) is the part of the cylinder \( x^2 + y^2 = 16 \) between \( z = 0 \) and \( z = h \).

Answer:
Parametrize the cylinder as \( x = 4 \cos \theta \) and \( y = 4 \sin \theta \) while \( z = t \) and the parameter domain is just \([0, 2\pi] \times [0, h]\). Then the integral to evaluate would be
\[
\int_0^{2\pi} \int_0^h (1 + 4 \cos \theta) 4 \, dt \, d\theta = 8h\pi.
\]

Note how \( 4 \cos \theta \) was substituted for \( x \) and the area element is \( 4 \, dt \, d\theta \).

13. Evaluate \( \int \int_S (1 + x) \, dA \) where \( S \) is the hemisphere \( x^2 + y^2 + z^2 = 16 \) between \( x = 0 \) and \( x = 4 \).

Answer:
Parametrize the sphere as \( x = 4 \sin \phi \cos \theta, y = 4 \sin \phi \sin \theta, \) and \( z = 4 \cos \phi \) and consider the values of the parameters. Since it is referred to as a hemisphere and involves \( x > 0, \theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \) and \( \phi \in [0, \pi] \). Then the area element is \( \sqrt{a^4 \sin \phi} \, d\theta \, d\phi \) and so the integral to evaluate is
\[
\int_0^{2\pi} \int_0^{\pi/2} (1 + 4 \sin \phi \cos \theta) 16 \sin \phi \, d\theta \, d\phi = 96\pi.
\]

14. For \((\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]\), let
\[
f(\theta, \alpha) \equiv (\cos \theta (2 + \cos \alpha), -\sin \theta (2 + \cos \alpha), \sin \alpha)^T.
\]

Find the area of \( f ([0, 2\pi] \times [0, 2\pi]) \).

Answer:
\[
Df(\theta, \alpha) = \begin{pmatrix}
-\sin(\theta)(2 + \cos \alpha) & -\cos \theta \sin \alpha \\
-\cos(\theta)(2 + \cos \alpha) & \sin \theta \sin \alpha \\
0 & \sin \alpha
\end{pmatrix}
\]
and so the area element is
\[
\det(Df^T Df)^{1/2} \, d\theta \, d\alpha = (4 + 4 \cos \alpha + \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha.
\]

Therefore, the area is
\[
\int_0^{2\pi} \int_0^{2\pi} (4 + 4 \cos \alpha + \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha = \int_0^{2\pi} \int_0^{2\pi} (2 + \cos \alpha) \, d\theta \, d\alpha = 8\pi^2.
\]
15. For \((\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]\), let

\[
f(\theta, \alpha) \equiv (\cos \theta (4 + 2 \cos \alpha), -\sin \theta (4 + 2 \cos \alpha), 2 \sin \alpha)^T.
\]

Also let \(h(x) = \cos \alpha\) where \(\alpha\) is such that

\[
x = (\cos \theta (4 + 2 \cos \alpha), -\sin \theta (4 + 2 \cos \alpha), 2 \sin \alpha)^T.
\]

Find \(\int_{f([0,2\pi] \times [0,2\pi])} h \, dV\).

Answer:

\[
Df(\theta, \alpha) = \begin{pmatrix}
-\sin (\theta) (4 + 2 \cos \alpha) & -2 \cos \theta \sin \alpha \\
-\cos (\theta) (4 + 2 \cos \alpha) & 2 \sin \theta \sin \alpha \\
0 & 2 \cos \alpha
\end{pmatrix}
\]

and so the area element is

\[
\det (Df^T Df)^{1/2} \, d\theta \, d\alpha = (64 + 64 \cos \alpha + 16 \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha.
\]

Therefore, the desired integral is

\[
\int_0^{2\pi} \int_0^{2\pi} (\cos \alpha) (64 + 64 \cos \alpha + 16 \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha
= \int_0^{2\pi} \int_0^{2\pi} (8 + 4 \cos \alpha) \, d\theta \, d\alpha = 8\pi^2
\]

16. For \((\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]\), let

\[
f(\theta, \alpha) \equiv (\cos \theta (3 + \cos \alpha), -\sin \theta (3 + \cos \alpha), \sin \alpha)^T.
\]

Also let \(h(x) = \cos^2 \theta\) where \(\theta\) is such that

\[
x = (\cos \theta (3 + \cos \alpha), -\sin \theta (3 + \cos \alpha), \sin \alpha)^T.
\]

Find \(\int_{f([0,2\pi] \times [0,2\pi])} h \, dV\).

Answer:

\[
Df(\theta, \alpha) = \begin{pmatrix}
-\sin (\theta) (3 + \cos \alpha) & -\cos \theta \sin \alpha \\
-\cos (\theta) (3 + \cos \alpha) & \sin \theta \sin \alpha \\
0 & \cos \alpha
\end{pmatrix}
\]

and so the area element is

\[
\det (Df^T Df)^{1/2} \, d\theta \, d\alpha = (9 + 6 \cos \alpha + \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha.
\]

Therefore, the desired integral is

\[
\int_0^{2\pi} \int_0^{2\pi} (\cos^2 \theta) (9 + 6 \cos \alpha + \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha
= \int_0^{2\pi} \int_0^{2\pi} (3 + \cos \alpha) \, d\theta \, d\alpha = 6\pi^2
\]

17. For \((\theta, \alpha) \in [0, 25] \times [0, 2\pi]\), let

\[
f(\theta, \alpha) \equiv (\cos \theta (4 + 2 \cos \alpha), -\sin \theta (4 + 2 \cos \alpha), 2 \sin \alpha + \theta)^T.
\]
Find a double integral which gives the area of $f ([0, 25] \times [0, 2\pi])$.

Answer:

In this case, $Df (\theta, \alpha) = 
\begin{pmatrix}
- \sin (\theta) (4 + 2 \cos \alpha) & -2 \cos \theta \sin \alpha \\
- \cos (\theta) (4 + 2 \cos \alpha) & 2 \sin \theta \sin \alpha \\
1 & 2 \cos \alpha
\end{pmatrix}
$ and so the area element is

$$
\det (Df^T Df) \, d\theta \, d\alpha = (68 + 64 \cos \alpha + 12 \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha
$$

and so the area is

$$
\int_0^{2\pi} \int_0^{2\pi} (68 + 64 \cos \alpha + 12 \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha.
$$

18. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, and $\beta$ a fixed real number, define $f (\theta, \alpha) \equiv 
\begin{pmatrix}
\cos \theta (2 + \cos \alpha), & - \cos \beta \sin \theta (2 + \cos \alpha), & \sin \beta \sin \theta (2 + \cos \alpha) + \cos \beta \sin \alpha, \\
- \cos \beta \sin \theta (2 + \cos \alpha) + \cos \beta \sin \alpha, & \sin \beta \sin \theta (2 + \cos \alpha) + \cos \beta \sin \alpha, & \cos \beta \sin \theta (2 + \cos \alpha) + \cos \beta \sin \alpha
\end{pmatrix}^T.

Find a double integral which gives the area of $f ([0, 2\pi] \times [0, 2\pi])$.

Answer:

$$
Df = 
\begin{pmatrix}
- \sin (2\theta + \theta \cos \alpha) & - \sin \alpha \cos \theta \\
-2 \cos \beta \sin \theta - \cos \beta \cos \theta \cos \alpha & \cos \beta \sin \theta \sin \alpha + \sin \beta \cos \alpha \\
2 \sin \beta \cos \theta + \sin \beta \cos \theta \cos \alpha & - \sin (\beta \sin \theta \sin \alpha) + \cos \beta \cos \alpha
\end{pmatrix}
$$

and so after many computations, the area element is $(4 + 4 \cos \alpha + \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha$. Therefore, the area is

$$
\int_0^{2\pi} \int_0^{2\pi} (2 + \cos \alpha) \, d\theta \, d\alpha = 8\pi^2.
$$

22.4 Exercises

1. Find a parameterization for the intersection of the planes $4x + 2y + 4z = 3$ and $6x - 2y = -1$.

2. Find a parameterization for the intersection of the plane $3x + y + z = 1$ and the circular cylinder $x^2 + y^2 = 1$.

3. Find a parameterization for the intersection of the plane $3x + 2y + 4z = 4$ and the elliptic cylinder $x^2 + 4z^2 = 16$.

4. Find a parameterization for the straight line joining $(1, 3, 1)$ and $(-2, 5, 3)$.

5. Find a parameterization for the intersection of the surfaces $4y + 3z = 3x^2 + 2$ and $3y + 2z = -x + 3$.

6. Find the area of $S$ if $S$ is the part of the circular cylinder $x^2 + y^2 = 4$ which lies between $z = 0$ and $z = 2 + y$.

7. Find the area of $S$ if $S$ is the part of the cone $x^2 + y^2 = 16z^2$ between $z = 0$ and $z = h$.

8. Parametrizing the cylinder $x^2 + y^2 = a^2$ by $x = a \cos v, y = a \sin v, z = u$, show that the area element is $dA = a \, du \, dv$.

9. Find the area enclosed by the limacon $r = 2 + \cos \theta$.

10. Find the surface area of the paraboloid $z = h (1 - x^2 - y^2)$ between $z = 0$ and $z = h$.

11. Evaluate $\int \int_S (1 + x) \, dA$ where $S$ is the part of the plane $4x + y + 3z = 12$ which is in the first octant.
12. Evaluate $\int \int_S (1 + x) \, dA$ where $S$ is the part of the cylinder $x^2 + y^2 = 9$ between $z = 0$ and $z = h$.

13. Evaluate $\int \int_S (1 + x) \, dA$ where $S$ is the hemisphere $x^2 + y^2 + z^2 = 4$ between $x = 0$ and $x = 2$.

14. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, let $f(\theta, \alpha) \equiv (\cos \theta (4 + \cos \alpha), -\sin \theta (4 + \cos \alpha), \sin \alpha)^T$.
Find the area of $f([0, 2\pi] \times [0, 2\pi])$.

15. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, let $f(\theta, \alpha) \equiv (\cos \theta (3 + 2 \cos \alpha), -\sin \theta (3 + 2 \cos \alpha), 2 \sin \alpha)^T$.
Also let $h(x) = \cos \alpha$ where $\alpha$ is such that $x = (\cos \theta (3 + 2 \cos \alpha), -\sin \theta (3 + 2 \cos \alpha), 2 \sin \alpha)^T$.
Find $\int f([0, 2\pi] \times [0, 2\pi]) h \, dV$.

16. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, let $f(\theta, \alpha) \equiv (\cos \theta (4 + 3 \cos \alpha), -\sin \theta (4 + 3 \cos \alpha), 3 \sin \alpha)^T$.
Also let $h(x) = \cos^2 \theta$ where $\theta$ is such that $x = (\cos \theta (4 + 3 \cos \alpha), -\sin \theta (4 + 3 \cos \alpha), 3 \sin \alpha)^T$.
Find $\int f([0, 2\pi] \times [0, 2\pi]) h \, dV$.

17. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, let $f(\theta, \alpha) \equiv (\cos \theta (4 + 2 \cos \alpha), -\sin \theta (4 + 2 \cos \alpha), 2 \sin \alpha + \theta)^T$.
Find a double integral which gives the area of $f([0, 2\pi] \times [0, 2\pi])$.

18. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, and $\beta$ a fixed real number, define $f(\theta, \alpha) \equiv \left(\frac{\cos \theta (3 + 2 \cos \alpha), -\cos \beta \sin \theta (3 + 2 \cos \alpha) + 2 \sin \beta \sin \alpha, \sin \beta \sin \theta (3 + 2 \cos \alpha) + 2 \cos \beta \sin \alpha}{\partial (u_1, u_2)}\right)^T$.
Find a double integral which gives the area of $f([0, 2\pi] \times [0, 2\pi])$.

19. In the case where $f : U \rightarrow \mathbb{R}^3$, show that
\[
\frac{\partial (x, y, z)}{\partial (u_1, u_2)} = |f_{u_1} \times f_{u_2}|.
\]
Thus the area element is $|f_{u_1} \times f_{u_2}| \, du_1 \, du_2$.

20. In spherical coordinates, $\phi = c, \rho \in [0, R]$ determines a cone. Find the area of this cone without doing any work involving Jacobians and such.
Calculus Of Vector Fields

23.0.1 Outcomes

1. Define and evaluate the divergence of a vector field in terms of Cartesian coordinates.
2. Define and evaluate the Curl of a vector field in Cartesian coordinates.
3. Discover vector identities involving the gradient, divergence, and curl.
4. Recall and verify the divergence theorem.
5. Apply the divergence theorem.

23.1 Divergence And Curl Of A Vector Field

Here the important concepts of divergence and curl are defined.

Definition 23.1.1 Let \( f : U \rightarrow \mathbb{R}^p \) for \( U \subseteq \mathbb{R}^p \) denote a vector field. A scalar valued function is called a scalar field. The function, \( f \) is called a \( C^k \) vector field if the function, \( f \) is a \( C^k \) function. For a \( C^1 \) vector field, as just described \( \nabla \cdot f(x) \equiv \text{div} f(x) \) known as the divergence, is defined as

\[
\nabla \cdot f(x) \equiv \text{div} f(x) \equiv \sum_{i=1}^{p} \frac{\partial f_i}{\partial x_i}(x).
\]

Using the repeated summation convention, this is often written as

\[
f_{i,i}(x) \equiv \partial_i f_i(x)
\]

where the comma indicates a partial derivative is being taken with respect to the \( i \)th variable and \( \partial_i \) denotes differentiation with respect to the \( i \)th variable. In words, the divergence is the sum of the \( i \)th derivative of the \( i \)th component function of \( f \) for all values of \( i \). If \( p = 3 \), the curl of the vector field yields another vector field and it is defined as follows.

\[
(\text{curl } (f)(x))_i \equiv (\nabla \times f(x))_i \equiv \varepsilon_{ijk} \partial_j f_k(x)
\]

where here \( \partial_j \) means the partial derivative with respect to \( x_j \) and the subscript of \( i \) in \( (\text{curl } (f)(x))_i \) means the \( i \)th Cartesian component of the vector, \( \text{curl } (f)(x) \). Thus the curl is evaluated by expanding the following determinant along the top row.

\[
\begin{vmatrix}
\frac{\partial}{\partial z} & \frac{\partial}{\partial y} & \frac{\partial}{\partial x} \\
f_1(x,y,z) & f_2(x,y,z) & f_3(x,y,z)
\end{vmatrix}
\]
Note the similarity with the cross product. Sometimes the curl is called rot. (Short for rotation not decay.) Also
\[ \nabla^2 f \equiv \nabla \cdot (\nabla f). \]
This last symbol is important enough that it is given a name, the Laplacian. It is also denoted by \( \Delta \). Thus \( \nabla^2 f = \Delta f \). In addition for \( f \) a vector field, the symbol \( f \cdot \nabla \) is defined as a “differential operator” in the following way.

\[ f \cdot \nabla (g) = f_1(x) \frac{\partial g(x)}{\partial x_1} + f_2(x) \frac{\partial g(x)}{\partial x_2} + \cdots + f_p(x) \frac{\partial g(x)}{\partial x_p}. \]

Thus \( f \cdot \nabla \) takes vector fields and makes them into new vector fields.

This definition is in terms of a given coordinate system but later coordinate free definitions of the curl and div are presented. For now, everything is defined in terms of a given Cartesian coordinate system. The divergence and curl have profound physical significance and this will be discussed later. For now it is important to understand their definition in terms of coordinates. Be sure you understand that for \( f \) a vector field, \( \text{div} \ f \) is a scalar field meaning it is a scalar valued function of three variables. For a scalar field, \( f \), \( \nabla f \) is a vector field described earlier on Page 336. For \( f \) a vector field having values in \( \mathbb{R}^3 \), \( \text{curl} \ f \) is a vector field.

**Example 23.1.2** Let \( f(x) = xy + (z - y)j + (\sin(x) + z)k \). Find \( \text{div} \ f \) and \( \text{curl} \ f \).

First the divergence of \( f \) is

\[ \frac{\partial (xy)}{\partial x} + \frac{\partial (z - y)}{\partial y} + \frac{\partial (\sin(x) + z)}{\partial z} = y + (-1) + 1 = y. \]

Now \( \text{curl} \ f \) is obtained by evaluating

\[ \begin{vmatrix} i & j & k \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ xy & z - y & \sin(x) + z \end{vmatrix} = \\
\begin{aligned} &i \left( \frac{\partial}{\partial y} (\sin(x) + z) - \frac{\partial}{\partial z} (z - y) \right) - j \left( \frac{\partial}{\partial x} (\sin(x) + z) - \frac{\partial}{\partial z} (xy) \right) + \\
&k \left( \frac{\partial}{\partial x} (z - y) - \frac{\partial}{\partial y} (xy) \right) = -i - \cos(x)j - xk. \end{aligned} \]

**23.1.1 Vector Identities**

There are many interesting identities which relate the gradient, divergence and curl.

**Theorem 23.1.3** Assuming \( f, g \) are a \( C^2 \) vector fields whenever necessary, the following identities are valid.

1. \( \nabla \cdot (\nabla \times f) = 0 \)
2. \( \nabla \times \nabla \phi = 0 \)
3. \( \nabla \times (\nabla \times f) = \nabla (\nabla \cdot f) - \nabla^2 f \) where \( \nabla^2 f \) is a vector field whose \( i \)th component is \( \nabla^2 f_i \).
4. \( \nabla \cdot (f \times g) = g \cdot (\nabla \times f) - f \cdot (\nabla \times g) \)
5. \( \nabla \times (f \times g) = (\nabla \cdot g) f - (\nabla \cdot f) g + (g \cdot \nabla) f - (f \cdot \nabla) g \)

**Proof:** These are all easy to establish if you use the repeated index summation convention and the reduction identities discussed on Page 153.

\[
\nabla \cdot (\nabla \times f) = \partial_i (\nabla \times f)_i = \partial_i (\varepsilon_{ijk} \partial_j f_k) = \varepsilon_{ijk} \partial_i (\partial_j f_k) = \varepsilon_{jik} \partial_j (\partial_i f_k) = -\varepsilon_{ijk} \partial_j (\partial_i f_k) = -\nabla \cdot (\nabla \times f).
\]

This establishes the first formula. The second formula is done similarly. Now consider the third.

\[
(\nabla \times (\nabla \times f))_i = \varepsilon_{ijk} \partial_j (\nabla \times f)_k = \varepsilon_{ijk} \partial_j (\varepsilon_{krs} \partial_r f_s) = \varepsilon_{kj} \varepsilon_{krs} \partial_j (\partial_r f_s) = (\delta_r^i \delta_j^s - \delta_i^s \delta_j^r) \partial_j (\partial_r f_s) = \partial_j (\partial_i f_j) - \partial_j (\partial_j f_i) = \partial_i (\partial_j f_j) - \partial_j (\partial_j f_i) = (\nabla (\nabla \cdot f) - \nabla^2 f)_i
\]

This establishes the third identity.

Consider the fourth identity.

\[
\nabla \cdot (f \times g) = \partial_i (f \times g)_i = \partial_i \varepsilon_{ijk} f_j g_k = \varepsilon_{ijk} \partial_i \varepsilon_{krs} f_r g_s = \varepsilon_{kij} \varepsilon_{krs} \partial_i (f_r g_s) = (\delta_r^i \delta_j^s - \delta_i^s \delta_j^r) \partial_j (f_r g_s) = \partial_j (f_i g_j) - \partial_j (f_j g_i) = (\partial_j f_i) f_j + g_j \partial_j f_i - (\partial_j f_j) g_i - f_j (\partial_j g_i) = ((\nabla \cdot g) f + (g \cdot \nabla) f - (\nabla \cdot g) f - (f \cdot \nabla) g)_i
\]

This proves the fourth identity.

Consider the fifth.

\[
(\nabla \times (f \times g))_i = \varepsilon_{ijk} \partial_j (f \times g)_k = \varepsilon_{ijk} \partial_j \varepsilon_{krs} f_r g_s = \varepsilon_{kij} \varepsilon_{krs} \partial_j (f_r g_s) = (\delta_r^i \delta_j^s - \delta_i^s \delta_j^r) \partial_j (f_r g_s) = \partial_j (f_i g_j) - \partial_j (f_j g_i) = (\partial_j f_i) f_j + g_j \partial_j f_i - (\partial_j f_j) g_i - f_j (\partial_j g_i) = (\nabla \cdot f + (f \cdot \nabla) f - (\nabla \cdot f) f - (f \cdot \nabla) g)_i
\]

and this establishes the fifth identity.

I think the important thing about the above is not that these identities can be proved and are valid as much as the method by which they were proved. The reduction identities on Page
153 were used to discover the identities. There is a difference between proving something
someone tells you about and both discovering what should be proved and proving it. This
notation and the reduction identity make the discovery of vector identities fairly routine
and this is why these things are of great significance.

**23.1.2 Vector Potentials**

One of the above identities says \( \nabla \cdot (\nabla \times \mathbf{f}) = 0 \). Suppose now \( \nabla \cdot \mathbf{g} = 0 \). Does it follow that
there exists \( \mathbf{f} \) such that \( \mathbf{g} = \nabla \times \mathbf{f} \)? It turns out that this is usually the case and when such
an \( \mathbf{f} \) exists, it is called a vector potential. Here is one way to do it, assuming everything is
defined so the following formulas make sense.

\[
\mathbf{f} (x, y, z) = \left( \int_0^z g_2 (x, y, t) \, dt, - \int_0^z g_1 (x, y, t) \, dt + \int_0^x g_3 (t, y, 0) \, dt, 0 \right)^T. \tag{23.1}
\]

In verifying this you need to use the following manipulation which will generally hold under
reasonable conditions but which has not been carefully shown yet.

\[
\frac{\partial}{\partial x} \int_a^b h (x, t) \, dt = \int_a^b \frac{\partial h}{\partial x} (x, t) \, dt. \tag{23.2}
\]

The above formula seems plausible because the integral is a sort of a sum and the derivative
of a sum is the sum of the derivatives. However, this sort of sloppy reasoning will get you
into all sorts of trouble. The formula involves the interchange of two limit operations, the
integral and the limit of a difference quotient. Such an interchange can only be accomplished
through a theorem. The following gives the necessary result. This lemma is stated without
proof.

**Lemma 23.1.4** Suppose \( h \) and \( \frac{\partial h}{\partial x} \) are continuous on the rectangle \( R = [c, d] \times [a, b] \). Then
(23.2) holds.

The second formula of Theorem 23.1.3 states \( \nabla \times \nabla \phi = 0 \). This suggests the following
question: Suppose \( \nabla \times \mathbf{f} = 0 \), does it follow there exists \( \phi \), a scalar field such that \( \nabla \phi = \mathbf{f} \)?
The answer to this is often yes and a theorem will be given and proved after the presentation
of Stoke’s theorem. This scalar field, \( \phi \), is called a scalar potential for \( \mathbf{f} \).

**23.1.3 The Weak Maximum Principle**

There is also a fundamental result having great significance which involves \( \nabla^2 \) called the
maximum principle. This principle says that if \( \nabla^2 u \geq 0 \) on a bounded open set, \( U \), then \( u \)
achieves its maximum value on the boundary of \( U \).

**Theorem 23.1.5** Let \( U \) be a bounded open set in \( \mathbb{R}^n \) and suppose \( u \in C^2 (U) \cap C (\overline{U}) \)
such that \( \nabla^2 u \geq 0 \) in \( U \). Then letting \( \partial U = \overline{U} \setminus U \), it follows that \( \max \{ u (x) : x \in \partial U \} = \max \{ u (x) : x \in \partial U \} \).

**Proof:** If this is not so, there exists \( \mathbf{x}_0 \in U \) such that \( u (\mathbf{x}_0) > \max \{ u (x) : x \in \partial U \} \equiv M \). Since \( U \) is bounded, there exists \( \varepsilon > 0 \) such that

\[
u (\mathbf{x}_0) > \max \{ u (x) + \varepsilon |x|^2 : x \in \partial U \} .
\]

Therefore, \( u (x) + \varepsilon |x|^2 \) also has its maximum in \( U \) because for \( \varepsilon \) small enough,

\[
u (\mathbf{x}_0) + \varepsilon |\mathbf{x}_0|^2 > u (\mathbf{x}_0) > \max \{ u (x) + \varepsilon |x|^2 : x \in \partial U \}
\]
for all \( x \in \partial U \).

Now let \( x_1 \) be the point in \( U \) at which \( u(x) + \varepsilon |x|^2 \) achieves its maximum. As an exercise you should show that \( \nabla^2 (f + g) = \nabla^2 f + \nabla^2 g \) and therefore, \( \nabla^2 \left( u(x) + \varepsilon |x|^2 \right) = \nabla^2 u(x) + 2n\varepsilon \). (Why?) Therefore,

\[
0 \geq \nabla^2 u(x_1) + 2n\varepsilon \geq 2n\varepsilon,
\]
a contradiction. This proves the theorem.

## 23.2 Exercises

1. Find \( \text{div} f \) and \( \text{curl} f \) where \( f \) is
   
   (a) \( (xyz, x^2 + \ln(xy), \sin x^2 + z)^T \)
   
   (b) \( (\sin x, \sin y, \sin z)^T \)
   
   (c) \( (f(x), g(y), h(z))^T \)
   
   (d) \( (x-2, y-3, z-6)^T \)
   
   (e) \( (y^2, 2xy, \cos z)^T \)
   
   (f) \( (f(y, z), g(x, z), h(y, z))^T \)

2. Prove formula 2 of Theorem 23.1.3.

3. Show that if \( u \) and \( v \) are \( C^2 \) functions, then \( \text{curl} (u\nabla v) = \nabla u \times \nabla v \).

4. Simplify the expression \( f \times (\nabla \times g) + g \times (\nabla \times f) + (f \cdot \nabla) g + (g \cdot \nabla) f \).

5. Simplify \( \nabla \times (v \times r) \) where \( r = (x, y, z)^T = xi + yj + zk \) and \( v \) is a constant vector.

6. Discover a formula which simplifies \( \nabla \cdot (v \nabla u) \).

7. Verify that \( \nabla \cdot (u\nabla v) - \nabla \cdot (v\nabla u) = u\nabla^2 v - v\nabla^2 u \).

8. Verify that \( \nabla^2 (uv) = v\nabla^2 u + 2(\nabla u \cdot \nabla v) + u\nabla^2 v \).

9. Functions, \( u \), which satisfy \( \nabla^2 u = 0 \) are called harmonic functions. Show the following functions are harmonic where ever they are defined.
   
   (a) \( 2xy \)
   
   (b) \( x^2 - y^2 \)
   
   (c) \( \sin x \cosh y \)
   
   (d) \( \ln (x^2 + y^2) \)
   
   (e) \( 1/\sqrt{x^2 + y^2 + z^2} \)

10. Verify the formula given in (23.1) is a vector potential for \( g \) assuming that \( \text{div} \ g = 0 \).

11. Show that if \( \nabla^2 u_k = 0 \) for each \( k = 1, 2, \ldots, m \), and \( c_k \) is a constant, then \( \nabla^2 \left( \sum_{k=1}^{m} c_k u_k \right) = 0 \) also.

12. In Theorem 23.1.5 why is \( \nabla^2 \left( \varepsilon |x|^2 \right) = 2n\varepsilon \)?
13. Using Theorem 23.1.5 prove the following: Let $f \in C(\partial U)$ ($f$ is continuous on $\partial U$) where $U$ is a bounded open set. Then there exists at most one solution, $u \in C^2(U) \cap C(\overline{U})$ and $\nabla^2 u = 0$ in $U$ with $u = f$ on $\partial U$. **Hint:** Suppose there are two solutions, $u_i$, $i = 1, 2$ and let $w = u_1 - u_2$. Then use the maximum principle.

14. Suppose $\mathbf{B}$ is a vector field and $\nabla \times \mathbf{A} = \mathbf{B}$. Thus $\mathbf{A}$ is a vector potential for $\mathbf{B}$. Show that $\mathbf{A} + \nabla \phi$ is also a vector potential for $\mathbf{B}$. Here $\phi$ is just a $C^2$ scalar field. Thus the vector potential is not unique.

### 23.3 The Divergence Theorem

The divergence theorem relates an integral over a set to one on the boundary of the set. It is also called Gauss’s theorem.

**Definition 23.3.1** A subset, $V$ of $\mathbb{R}^3$ is called cylindrical in the $x$ direction if it is of the form

$$V = \{(x, y, z) : \phi(y, z) \leq x \leq \psi(y, z) \text{ for } (y, z) \in D\}$$

where $D$ is a subset of the $yz$ plane. $V$ is cylindrical in the $z$ direction if

$$V = \{(x, y, z) : \phi(x, y) \leq z \leq \psi(x, y) \text{ for } (x, y) \in D\}$$

where $D$ is a subset of the $xy$ plane, and $V$ is cylindrical in the $y$ direction if

$$V = \{(x, y, z) : \phi(x, z) \leq y \leq \psi(x, z) \text{ for } (x, z) \in D\}$$

where $D$ is a subset of the $xz$ plane. If $V$ is cylindrical in the $z$ direction, denote by $\partial V$ the boundary of $V$ defined to be the points of the form $(x, y, \phi(x, y)), (x, y, \psi(x, y))$ for $(x, y) \in D$, along with points of the form $(x, y, z)$ where $(x, y) \in \partial D$ and $\phi(x, y) \leq z \leq \psi(x, y)$. Points on $\partial D$ are defined to be those for which every open ball contains points which are in $D$ as well as points which are not in $D$. A similar definition holds for $\partial V$ in the case that $V$ is cylindrical in one of the other directions.

The following picture illustrates the above definition in the case of $V$ cylindrical in the $z$ direction.
Of course, many three dimensional sets are cylindrical in each of the coordinate directions. For example, a ball or a rectangle or a tetrahedron are all cylindrical in each direction. The following lemma allows the exchange of the volume integral of a partial derivative for an area integral in which the derivative is replaced with multiplication by an appropriate component of the unit exterior normal.

**Lemma 23.3.2** Suppose \( V \) is cylindrical in the \( z \) direction and that \( \phi \) and \( \psi \) are the functions in the above definition. Assume \( \phi \) and \( \psi \) are \( C^1 \) functions and suppose \( F \) is a \( C^1 \) function defined on \( V \). Also, let \( n = (n_x, n_y, n_z) \) be the unit exterior normal to \( \partial V \). Then

\[
\int \int \int_V \frac{\partial F}{\partial z}(x, y, z) \, dV = \int \int_{\partial V} F \, n_z \, dA.
\]

**Proof:** From the fundamental theorem of calculus,

\[
\int \int \int_V \frac{\partial F}{\partial z}(x, y, z) \, dV = \int \int_D \int_{\phi(x,y)}^{\psi(x,y)} \frac{\partial F}{\partial z}(x, y, z) \, dz \, dx \, dy \tag{23.3}\]

\[
= \int \int_D [F(x, y, \psi(x, y)) - F(x, y, \phi(x, y))] \, dx \, dy
\]

Now the unit exterior normal on the top of \( V \), the surface \( (x, y, \psi(x, y)) \) is

\[
\frac{1}{\sqrt{\psi_x^2 + \psi_y^2 + 1}} \left(-\psi_x, -\psi_y, 1\right).
\]

This follows from the observation that the top surface is the level surface, \( z = \psi(x, y) = 0 \) and so the gradient of this function of three variables is perpendicular to the level surface. It points in the correct direction because the \( z \) component is positive. Therefore, on the top surface,

\[
n_z = \frac{1}{\sqrt{\psi_x^2 + \psi_y^2 + 1}}
\]
Similarly, the unit normal to the surface on the bottom is
\[ n_z = \frac{-1}{\sqrt{\phi_x^2 + \phi_y^2 + 1}} \]
and so on the bottom surface,
\[ n_z = \frac{-1}{\sqrt{\phi_x^2 + \phi_y^2 + 1}} \]
Note that here the \( z \) component is negative because since it is the outer normal it must point down. On the lateral surface, the one where \((x, y) \in \partial D\) and \( z \in [\phi(x, y), \psi(x, y)]\), \( n_z = 0 \).

The area element on the top surface is \( dA = \sqrt{\psi_x^2 + \psi_y^2 + 1} \, dx \, dy \) while the area element on the bottom surface is \( \sqrt{\phi_x^2 + \phi_y^2 + 1} \, dx \, dy \). Therefore, the last expression in (23.3) is of the form,
\[
\int \int_D F(x, y, \psi(x, y)) \left( \frac{n_z}{\sqrt{\psi_x^2 + \psi_y^2 + 1}} \right) dA + \int \int_{\text{Lateral surface}} fn_z \, dA,
\]
the last term equaling zero because on the lateral surface, \( n_z = 0 \). Therefore, this reduces to \( \int_{\partial V} fn_z \, dA \) as claimed.

The following corollary is entirely similar to the above.

**Corollary 23.3.3** If \( V \) is cylindrical in the \( y \) direction, then
\[
\int \int \int_V \frac{\partial F}{\partial y} \, dV = \int \int_{\partial V} F n_y \, dA
\]
and if \( V \) is cylindrical in the \( x \) direction, then
\[
\int \int \int_V \frac{\partial F}{\partial x} \, dV = \int \int_{\partial V} F n_x \, dA
\]
With this corollary, here is a proof of the divergence theorem.

**Theorem 23.3.4** Let \( V \) be cylindrical in each of the coordinate directions and let \( F \) be a \( C^1 \) vector field defined on \( V \). Then
\[
\int \int \int_V \nabla \cdot F \, dV = \int \int_{\partial V} F \cdot n \, dA.
\]
Proof: From the above lemma and corollary,

\[
\int \int \int \nabla \cdot \mathbf{F} \, dV = \int \int \int \left( \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z} \right) \, dV
\]

\[
= \int_{\partial V} (F_1 n_x + F_2 n_y + F_3 n_z) \, dA
\]

\[
= \int_{\partial V} \mathbf{F} \cdot \mathbf{n} \, dA.
\]

This proves the theorem.

The divergence theorem holds for much more general regions than this. Suppose for example you have a complicated region which is the union of finitely many disjoint regions of the sort just described which are cylindrical in each of the coordinate directions. Then the volume integral over the union of these would equal the sum of the integrals over the disjoint regions. If the boundaries of two of these regions intersect, then the area integrals will cancel out on the intersection because the unit exterior normals will point in opposite directions. Therefore, the sum of the integrals over the boundaries of these disjoint regions will reduce to an integral over the boundary of the union of these. Hence the divergence theorem will continue to hold. For example, consider the following picture. If the divergence theorem holds for each \( V_i \) in the following picture, then it holds for the union of these two.

\[\begin{array}{c}
\includegraphics[width=0.2\textwidth]{figure.png}
\end{array}\]

General formulations of the divergence theorem involve Hausdorff measures and the Lebesgue integral, a better integral than the old fashioned Riemann integral which has been obsolete now for almost 100 years. When all is said and done, one finds that the conclusion of the divergence theorem is usually true and the theorem can be used with confidence.

### 23.3.1 Coordinate Free Concept Of Divergence

The divergence theorem also makes possible a coordinate free definition of the divergence.

**Theorem 23.3.5** Let \( B(x, \delta) \) be the ball centered at \( x \) having radius \( \delta \) and let \( \mathbf{F} \) be a \( C^1 \) vector field. Then letting \( v(B(x, \delta)) \) denote the volume of \( B(x, \delta) \) given by

\[
\int_{B(x,\delta)} \, dV,
\]

it follows

\[
\text{div} \mathbf{F}(x) = \lim_{\delta \to 0^+} \frac{1}{v(B(x, \delta))} \int_{\partial B(x, \delta)} \mathbf{F} \cdot \mathbf{n} \, dA.
\] (23.4)

**Proof:** The divergence theorem holds for balls because they are cylindrical in every direction. Therefore,

\[
\frac{1}{v(B(x, \delta))} \int_{\partial B(x, \delta)} \mathbf{F} \cdot \mathbf{n} \, dA = \frac{1}{v(B(x, \delta))} \int_{B(x, \delta)} \text{div} \mathbf{F}(y) \, dV.
\]
Therefore, since \( \text{div} \, \mathbf{F}(x) \) is a constant,

\[
\left| \text{div} \, \mathbf{F}(x) - \frac{1}{v(B(x, \delta))} \int_{\partial B(x, \delta)} \mathbf{F} \cdot \mathbf{n} \, dA \right|
\]

\[
= \left| \text{div} \, \mathbf{F}(x) - \frac{1}{v(B(x, \delta))} \int_{B(x, \delta)} \text{div} \, \mathbf{F}(y) \, dV \right|
\]

\[
= \frac{1}{v(B(x, \delta))} \int_{B(x, \delta)} (\text{div} \, \mathbf{F}(x) - \text{div} \, \mathbf{F}(y)) \, dV
\]

\[
\leq \frac{1}{v(B(x, \delta))} \int_{B(x, \delta)} |\text{div} \, \mathbf{F}(x) - \text{div} \, \mathbf{F}(y)| \, dV
\]

\[
\leq \frac{1}{v(B(x, \delta))} \int_{B(x, \delta)} \frac{\varepsilon}{2} \, dV < \varepsilon
\]

whenever \( \varepsilon \) is small enough due to the continuity of \( \text{div} \, \mathbf{F} \). Since \( \varepsilon \) is arbitrary, this shows (23.4).

How is this definition independent of coordinates? It only involves geometrical notions of volume and dot product. This is why. Imagine rotating the coordinate axes, keeping all distances the same and expressing everything in terms of the new coordinates. The divergence would still have the same value because of this theorem.

### 23.4 Some Applications Of The Divergence Theorem

#### 23.4.1 Hydrostatic Pressure

Imagine a fluid which does not move which is acted on by an acceleration, \( \mathbf{g} \). Of course the acceleration is usually the acceleration of gravity. Also let the density of the fluid be \( \rho \), a function of position. What can be said about the pressure, \( p \), in the fluid? Let \( B(x, \varepsilon) \) be a small ball centered at the point, \( x \). Then the force the fluid exerts on this ball would equal

\[
- \int_{\partial B(x, \varepsilon)} \rho \mathbf{n} \, dA.
\]

Here \( \mathbf{n} \) is the unit exterior normal at a small piece of \( \partial B(x, \varepsilon) \) having area \( dA \). By the divergence theorem, (see Problem 1 on Page 442) this integral equals

\[
- \int_{B(x, \varepsilon)} \nabla p \, dV.
\]

Also the force acting on this small ball of fluid is

\[
\int_{B(x, \varepsilon)} \rho \mathbf{g} \, dV.
\]

Since it is given that the fluid does not move, the sum of these forces must equal zero. Thus

\[
\int_{B(x, \varepsilon)} \rho \mathbf{g} \, dV = \int_{B(x, \varepsilon)} \nabla p \, dV.
\]
Since this must hold for any ball in the fluid of any radius, it must be that
\[ \nabla p = \rho g. \tag{23.5} \]

It turns out that the pressure in a lake at depth \( z \) is equal to 62.5\( z \). This is easy to see from (23.5). In this case, \( g = gk \) where \( g = 32 \text{ feet/sec}^2 \). The weight of a cubic foot of water is 62.5 pounds. Therefore, the mass in slugs of this water is 62.5/32. Since it is a cubic foot, this is also the density of the water in slugs per cubic foot. Also, it is normally assumed that water is incompressible\(^1\). Therefore, this is the mass of water at any depth. Therefore,
\[ \frac{\partial p}{\partial x} + \frac{\partial p}{\partial y} + \frac{\partial p}{\partial z} = \frac{62.5}{32} \times 32k. \]
and so \( p \) does not depend on \( x \) and \( y \) and is only a function of \( z \). It follows \( p(0) = 0 \), and \( p'(z) = 62.5 \). Therefore, \( p(x, y, z) = 62.5z \). This establishes the claim. This is interesting but (23.5) is more interesting because it does not require \( \rho \) to be constant.

### 23.4.2 Archimedes Law Of Buoyancy

Archimedes principle states that when a solid body is immersed in a fluid the net force acting on the body by the fluid is directly up and equals the total weight of the fluid displaced.

Denote the set of points in three dimensions occupied by the body as \( V \). Then for \( dA \) an increment of area on the surface of this body, the force acting on this increment of area would equal \(-p \, dA \, n \) where \( n \) is the exterior unit normal. Therefore, since the fluid does not move,
\[ \int_{\partial V} -pn \, dA = \int_{V} -\nabla p \, dV = \int_{V} \rho g \, dV \, k \]
Which equals the total weight of the displaced fluid and you note the force is directed upward as claimed. Here \( \rho \) is the density and (23.5) is being used. There is an interesting point in the above explanation. Why does the second equation hold? Imagine that \( V \) were filled with fluid. Then the equation follows from (23.5) because in this equation \( g = -gk \).

### 23.4.3 Equations Of Heat And Diffusion

Let \( x \) be a point in three dimensional space and let \((x_1, x_2, x_3)\) be Cartesian coordinates of this point. Let there be a three dimensional body having density, \( \rho = \rho(x, t) \).

The heat flux, \( J \), in the body is defined as a vector which has the following property.

The heat flux, \( J \), in the body is defined as a vector which has the following property.
\[ \text{Rate at which heat crosses } S = \int_{S} J \cdot n \, dA \]
where \( n \) is the unit normal in the desired direction. Thus if \( V \) is a three dimensional body,
\[ \text{Rate at which heat leaves } V = \int_{\partial V} J \cdot n \, dA \]
where \( n \) is the unit exterior normal.

Fourier’s law of heat conduction states that the heat flux, \( J \) satisfies \( J = -k \nabla (u) \) where \( u \) is the temperature and \( k = k(u, x, t) \) is called the coefficient of thermal conductivity. This changes depending on the material. It also can be shown by experiment to change

\(^1\)There is no such thing as an incompressible fluid but this doesn’t stop people from making this assumption.
with temperature. This equation for the heat flux states that the heat flows from hot places toward colder places in the direction of greatest rate of decrease in temperature. Let \( c(x, t) \) denote the specific heat of the material in the body. This means the amount of heat within \( V \) is given by the formula \( \int \int \int_V \rho(x, t) c(x, t) u(x, t) \, dV \). Suppose also there are sources for the heat within the material given by \( f(x, u, t) \). If \( f \) is positive, the heat is increasing while if \( f \) is negative the heat is decreasing. For example such sources could result from a chemical reaction taking place. Then the divergence theorem can be used to verify the following equation for \( u \). Such an equation is called a reaction diffusion equation.

\[
\frac{\partial}{\partial t} \left( \rho(x, t) c(x, t) u(x, t) \right) = \nabla \cdot (k(u, x, t) \nabla u(x, t)) + f(x, u, t). \tag{23.6}
\]

Take an arbitrary \( V \) for which the divergence theorem holds. Then the time rate of change of the heat in \( V \) is

\[
\frac{d}{dt} \int_V \rho(x, t) c(x, t) u(x, t) \, dV = \int_V \frac{\partial}{\partial t} \left( \rho(x, t) c(x, t) u(x, t) \right) \, dV
\]

where, as in the preceding example, this is a physical derivation so the consideration of hard mathematics is not necessary. Therefore, from the Fourier law of heat conduction,

\[
\frac{d}{dt} \int_V \rho(x, t) c(x, t) u(x, t) \, dV = \int_V \frac{\partial}{\partial t} \left( \rho(x, t) c(x, t) u(x, t) \right) \, dV
\]

\[
\int_{\partial V} \rho \nabla u \cdot n \, dA + \int_V f(x, u, t) \, dV = \int \int \int_V \left( \nabla \cdot (k \nabla u) + f \right) \, dV.
\]

Since this holds for every sample volume, \( V \) it must be the case that the above reaction diffusion equation, (23.6) holds. Note that more interesting equations can be obtained by letting more of the quantities in the equation depend on temperature. However, the above is a fairly hard equation and people usually assume the coefficient of thermal conductivity depends only on \( x \) and that the reaction term, \( f \) depends only on \( x \) and \( t \) and that \( \rho \) and \( c \) are constant. Then it reduces to the much easier equation,

\[
\frac{\partial}{\partial t} u(x, t) = \frac{1}{\rho c} \nabla \cdot (k(x) \nabla u(x, t)) + f(x, t).
\tag{23.7}
\]

This is often referred to as the heat equation. Sometimes there are modifications of this in which \( k \) is not just a scalar but a matrix to account for different heat flow properties in different directions. However, they are not much harder than the above. The major mathematical difficulties result from allowing \( k \) to depend on temperature.

It is known that the heat equation is not correct even if the thermal conductivity did not depend on \( u \) because it implies infinite speed of propagation of heat. However, this does not prevent people from using it.

### 23.4.4 Balance Of Mass

Let \( y \) be a point in three dimensional space and let \((y_1, y_2, y_3)\) be Cartesian coordinates of this point. Let \( V \) be a region in three dimensional space and suppose a fluid having density, \( \rho(y, t) \) and velocity, \( v(y, t) \) is flowing through this region. Then the mass of fluid leaving \( V \) per unit time is given by the area integral, \( \int_{\partial V} \rho(y, t) v(y, t) \cdot n \, dA \) while the total mass of the fluid enclosed in \( V \) at a given time is \( \int_V \rho(y, t) \, dV \). Also suppose mass originates at the
rate $f(y, t)$ per cubic unit per unit time within this fluid. Then the conclusion which can be drawn through the use of the divergence theorem is the following fundamental equation known as the mass balance equation.

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = f(y, t)$$  \hspace{1cm} (23.8)

To see this is so, take an arbitrary $V$ for which the divergence theorem holds. Then the time rate of change of the mass in $V$ is

$$\frac{\partial}{\partial t} \int_V \rho(y, t) \, dV = \int_V \frac{\partial \rho(y, t)}{\partial t} \, dV$$

where the derivative was taken under the integral sign with respect to $t$. (This is a physical derivation and therefore, it is not necessary to fuss with the hard mathematics related to the change of limit operations. You should expect this to be true under fairly general conditions because the integral is a sort of sum and the derivative of a sum is the sum of the derivatives.) Therefore, the rate of change of mass, $\frac{\partial}{\partial t} \int_V \rho(y, t) \, dV$, equals

$$\int_V \frac{\partial \rho(y, t)}{\partial t} \, dV = \int_{\partial V} \rho(y, t) \mathbf{v}(y, t) \cdot \mathbf{n} \, dA + \int_V f(y, t) \, dV$$

$$= \int_V \left( \nabla \cdot (\rho(y, t) \mathbf{v}(y, t)) + f(y, t) \right) \, dV.$$  

Since this holds for every sample volume, $V$ it must be the case that the equation of continuity holds. Again, there are interesting mathematical questions here which can be explored but since it is a physical derivation, it is not necessary to dwell too much on them. If all the functions involved are continuous, it is certainly true but it is true under far more general conditions than that.

Also note this equation applies to many situations and $f$ might depend on more than just $y$ and $t$. In particular, $f$ might depend also on temperature and the density, $\rho$. This would be the case for example if you were considering the mass of some chemical and $f$ represented a chemical reaction. Mass balance is a general sort of equation valid in many contexts.

### 23.4.5 Balance Of Momentum

This example is a little more substantial than the above. It concerns the balance of momentum for a continuum. To see a full description of all the physics involved, you should consult a book on continuum mechanics. The situation is of a material in three dimensions and it deforms and moves about in three dimensions. This means this material is not a rigid body. Let $B_0$ denote an open set identifying a chunk of this material at time $t = 0$ and let $B_t$ be an open set which identifies the same chunk of material at time $t > 0$.

Let $y(t, x) = (y_1(t, x), y_2(t, x), y_3(t, x))$ denote the position with respect to Cartesian coordinates at time $t$ of the point whose position at time $t = 0$ is $x = (x_1, x_2, x_3)$. The coordinates, $x$ are sometimes called the reference coordinates and sometimes the material coordinates and sometimes the Lagrangian coordinates. The coordinates, $y$ are called the Eulerian coordinates or sometimes the spacial coordinates and the function, $(t, x) \rightarrow y(t, x)$ is called the motion. Thus

$$y(0, x) = x.$$  \hspace{1cm} (23.9)

The derivative,

$$D_2 y(t, x)$$
CALCULUS OF VECTOR FIELDS

is called the deformation gradient. Recall the notation means you fix $t$ and consider the function, $x \to y(t, x)$, taking its derivative. Since it is a linear transformation, it is represented by the usual matrix, whose $ij^{th}$ entry is given by

$$F_{ij}(x) = \frac{\partial y_i(t, x)}{\partial x_j}.$$  

Let $\rho(t, y)$ denote the density of the material at time $t$ at the point, $y$ and let $\rho_0(x)$ denote the density of the material at the point, $x$. Thus $\rho_0(x) = \rho(0, x) = \rho(0, y(0, x))$. The first task is to consider the relationship between $\rho(t, y)$ and $\rho_0(x)$.

**Lemma 23.4.1** $\rho_0(x) = \rho(t, y(t, x)) \det(F)$ and in any reasonable physical motion, $\det(F) > 0$.

**Proof:** Let $V_0$ represent a small chunk of material at $t = 0$ and let $V_t$ represent the same chunk of material at time $t$. I will be a little sloppy and refer to $V_0$ as the small chunk of material at time $t = 0$ and $V_t$ as the chunk of material at time $t$ rather than an open set representing the chunk of material. Then by the change of variables formula for multiple integrals,

$$\int_{V_t} dV = \int_{V_0} |\det(F)| \, dV.$$

If $\det(F) = 0$ for some $t$ the above formula shows that the chunk of material went from positive volume to zero volume and this is not physically possible. Therefore, it is impossible that $\det(F)$ can equal zero. However, at $t = 0$, $F = I$, the identity because of (23.9).

Therefore, $\det(F) = 1$ at $t = 0$ and if it is assumed $t \to \det(F)$ is continuous it follows by the intermediate value theorem that $\det(F) > 0$ for all $t$. Of course it is not known for sure this function is continuous but the above shows why it is at least reasonable to expect $\det(F) > 0$.

Now using the change of variables formula,

$$\text{mass of } V_t = \int_{V_t} \rho(t, y) \, dV = \int_{V_0} \rho(t, y(t, x)) \det(F) \, dV = \text{mass of } V_0 = \int_{V_0} \rho_0(x) \, dV.$$  

Since $V_0$ is arbitrary, it follows $\rho_0(x) = \rho(t, y(t, x)) \det(F)$ as claimed. Note this shows that $\det(F)$ is a magnification factor for the density.

Now consider a small chunk of material, $B_t$ at time $t$ which corresponds to $B_0$ at time $t = 0$. The total linear momentum of this material at time $t$ is

$$\int_{B_t} \rho(t, y) \mathbf{v}(t, y) \, dV$$

where $\mathbf{v}$ is the velocity. By Newton’s second law, the time rate of change of this linear momentum should equal the total force acting on the chunk of material. In the following derivation, $dV(y)$ will indicate the integration is taking place with respect to the variable,
y. By Lemma 23.4.1 and the change of variables formula for multiple integrals
\[
\frac{d}{dt} \left( \int_{B_t} \rho(t, y) \mathbf{v}(t, y) \, dV(y) \right) = \frac{d}{dt} \left( \int_{B_0} \rho(t, y(t, x)) \mathbf{v}(t, y(t, x)) \, dV(x) \right) = \int_{B_0} \rho_0(x) \mathbf{v}(t, y(t, x)) \, dV(x)
\]
\[
= \int_{B_0} \rho_0(x) \left[ \frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial y_i} \frac{\partial y_i}{\partial t} \right] \, dV(x) = \int_{B_0} \rho(t, y) \left[ \frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial y_i} \frac{\partial y_i}{\partial t} \right] \det(F) \, dV(y)
\]
\[
= \int_{B_t} \rho(t, y) \left[ \frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial y_i} \frac{\partial y_i}{\partial t} \right] \, dV(y).
\]

Having taken the derivative of the total momentum, it is time to consider the total force acting on the chunk of material.

The force comes from two sources, a body force, \( \mathbf{b} \) and a force which act on the boundary of the chunk of material called a traction force. Typically, the body force is something like
\[
\text{the usual repeated index summation convention,}
\]
\[
\text{equate the time derivative of the total linear momentum with the applied forces and using}
\]
\[
\frac{\partial}{\partial t} \text{the term,}
\]
\[
\text{where here div } B \text{ is exchanged for the}
\]
\[
\text{matrix is symmetric, } T_{ij} = T_{ji}. \text{ It is called the Cauchy stress. Using Newton’s second law to equate the time derivative of the total linear momentum with the applied forces and using the usual repeated index summation convention,}
\]
\[
\int_{\partial B_t} \mathbf{s}(t, y, \mathbf{n}) \, dA
\]
where \( \mathbf{n} \) is the unit exterior normal. Thus the traction force depends on position, time, and the orientation of the boundary of \( B_t \). Cauchy showed the existence of a linear transformation, \( T(t, y) \) such that \( T(t, y) \mathbf{n} = \mathbf{s}(t, y, \mathbf{n}) \). It follows there is a matrix, \( T_{ij}(t, y) \) such that the \( i^{th} \) component of \( \mathbf{s} \) is given by \( s_i(t, y, \mathbf{n}) = T_{ij}(t, y) n_j \). Cauchy also showed this matrix is symmetric, \( T_{ij} = T_{ji} \). Thus you might see this written as
\[
\rho \mathbf{v} = \mathbf{b} + \text{div } (T).
\]
The above formulation of the balance of momentum involves the spatial coordinates, $y$, but people also like to formulate momentum balance in terms of the material coordinates, $x$. Of course this changes everything.

The momentum in terms of the material coordinates is

$$\int_{B_0} \rho_0(x) \mathbf{v}(t,x) \, dV$$

and so, since $x$ does not depend on $t$,

$$\frac{d}{dt} \left( \int_{B_0} \rho_0(x) \mathbf{v}(t,x) \, dV \right) = \int_{B_0} \rho_0(x) \mathbf{v}_t(t,x) \, dV.$$  

As indicated earlier, this is a physical derivation and so the mathematical questions related to interchange of limit operations are ignored. This must equal the total applied force. Thus

$$\int_{B_0} \rho_0(x) \mathbf{v}_t(t,x) \, dV = \int_{B_0} b_0(t,x) \, dV + \int_{\partial B_t} T_{ij} n_j dA,$$  

the first term on the right being the contribution of the body force given per unit volume in the material coordinates and the last term being the traction force discussed earlier. The task is to write this last integral as one over $\partial B_0$. For $y \in \partial B_t$ there is a unit outer normal, $\mathbf{n}$. Here $y = y(t,x)$ for $x \in \partial B_0$. Then define $\mathbf{N}$ to be the unit outer normal to $B_0$ at the point, $x$. Near the point $y \in \partial B_t$, the surface, $\partial B_t$ is given parametrically in the form $y = y(s,t)$ for $(s,t) \in D \subseteq \mathbb{R}^2$ and it can be assumed the unit normal to $\partial B_t$ near this point is

$$\mathbf{n} = \frac{y_s(s,t) \times y_t(s,t)}{|y_s(s,t) \times y_t(s,t)|}$$

with the area element given by $|y_s(s,t) \times y_t(s,t)| \, ds \, dt$. This is true for $y \in P_t \subseteq \partial B_t$, a small piece of $\partial B_t$. Therefore, the last integral in (23.10) is the sum of integrals over small pieces of the form

$$\int_{P_t} T_{ij} n_j dA$$  

(23.11)

where $P_t$ is parametrized by $y(s,t)$, $(s,t) \in D$. Thus the integral in (23.11) is of the form

$$\int_D T_{ij} (y(s,t)) (y_s(s,t) \times y_t(s,t)) \, ds \, dt.$$  

By the chain rule this equals

$$\int_D T_{ij} (y(s,t)) \left( \frac{\partial y}{\partial x_\alpha} \frac{\partial x_\alpha}{\partial s} \times \frac{\partial y}{\partial x_\beta} \frac{\partial x_\beta}{\partial t} \right) \, ds \, dt.$$  

Remember $y = y(t,x)$ and it is always assumed the mapping $x \rightarrow y(t,x)$ is one to one and so, since on the surface $\partial B_t$ near $y$, the points are functions of $(s,t)$, it follows $x$ is also a function of $(s,t)$. Now by the properties of the cross product, this last integral equals

$$\int_D T_{ij} (x(s,t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \left( \frac{\partial y}{\partial x_\alpha} \times \frac{\partial y}{\partial x_\beta} \right) \, ds \, dt$$  

(23.12)

where here $x(s,t)$ is the point of $\partial B_0$ which corresponds with $y(s,t) \in \partial B_t$. Thus $T_{ij} (x(s,t)) = T_{ij} (y(s,t))$. (Perhaps this is a slight abuse of notation because $T_{ij}$ is defined on $\partial B_t$, not on $\partial B_0$, but it avoids introducing extra symbols.) Next (23.12) equals

$$\int_D T_{ij} (x(s,t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \varepsilon_{jkl} \frac{\partial y_l}{\partial x_\alpha} \frac{\partial y_k}{\partial x_\beta} \, ds \, dt$$
23.4. SOME APPLICATIONS OF THE DIVERGENCE THEOREM

\[ = \int_D T_{ij} (x(s,t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \varepsilon_{cab} \delta_{jc} \] 
\[ = \int_D T_{ij} (x(s,t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \varepsilon_{cab} \delta_{jc} \] 
\[ = \int_D (\det F) T_{ij} (x(s,t)) \varepsilon_{p\alpha\beta} \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \frac{\partial y_c}{\partial x_p} ds dt. \]

Now \( \frac{\partial x_p}{\partial y_j} = F_j^{-1} \) and also
\[ \varepsilon_{p\alpha\beta} \] 
so the result just obtained is of the form
\[ \int_D (\det F) F_j^{-1} T_{ij} (x(s,t)) (x_s \times x_t)_p ds dt = \]
\[ \int_D (\det F) T_{ij} (x(s,t)) (F^{-T})_{jp} (x_s \times x_t)_p ds dt. \]

This has transformed the integral over \( P_t \) to one over \( P_0 \), the part of \( \partial B_0 \) which corresponds with \( P_t \). Thus the last integral is of the form
\[ \int_{P_0} (\det F) (F^{-T}T)_{ip} N_p dA \]

Summing these up over the pieces of \( \partial B_t \) and \( \partial B_0 \) yields the last integral in (23.10) equals
\[ \int_{\partial B_0} (\det F) (F^{-T}T)_{ip} N_p dA \]

and so the balance of momentum in terms of the material coordinates becomes
\[ \int_{B_0} \rho_0 (x) \mathbf{v}_t (t,x) \ dV = \int_{B_0} b_0 (t,x) \ dV + \int_{\partial B_0} (\det F) (F^{-T}T)_{ip} N_p dA \]

The matrix, \( (\det F) (F^{-T}T)_{ip} \) is called the Piola Kirchhoff stress, \( S \). An application of the divergence theorem yields
\[ \int_{B_0} \rho_0 (x) \mathbf{v}_t (t,x) \ dV = \int_{B_0} b_0 (t,x) \ dV + \int_{B_0} \frac{\partial (\det F) (F^{-T}T)_{ip}}{\partial x_p} \ dV. \]

Since \( B_0 \) is arbitrary, a balance law for momentum in terms of the material coordinates is obtained
\[ \rho_0 (x) \mathbf{v}_t (t,x) = b_0 (t,x) + \frac{\partial (\det F) (F^{-T}T)_{ip}}{\partial x_p} \]
\[ = b_0 (t,x) + \text{div} (\det F) (F^{-T}T) \]
\[ = b_0 (t,x) + \text{div} S. \quad (23.13) \]
The main purpose of this presentation is to show how the divergence theorem is used in a significant way to obtain balance laws and to indicate a very interesting direction for further study. To continue, one needs to specify \( T \) or \( S \) as an appropriate function of things related to the motion, \( y \). Often the thing related to the motion is something called the strain and such relationships between the stress and the strain are known as constitutive laws. The proper formulation of constitutive laws involves more physical considerations such as frame indifference in which it is required the response of the system cannot depend on the manner in which the Cartesian coordinate system was chosen. There are also many other physical properties which can be included and which require a certain form for the constitutive equations. These considerations are outside the scope of this book and require a considerable amount of linear algebra.

There are also balance laws for energy which you may study later but these are more problematic than the balance laws for mass and momentum. However, the divergence theorem is used in these also.

### 23.4.6 The Wave Equation

As an example of how the balance law of momentum is used to obtain an important equation of mathematical physics, suppose \( S = kF \) where \( k \) is a constant and \( F \) is the deformation gradient and let \( u = y - x \). Thus \( u \) is the displacement. Then from (23.13) you can verify the following holds.

\[
\rho_0 (x) u_{tt} (t, x) = b_0 (t, x) + k \Delta u (t, x)
\]

(23.14)

In the case where \( \rho_0 \) is a constant and \( b_0 = 0 \), this yields

\[
u_{tt} - c \Delta u = 0.
\]

The wave equation is \( u_{tt} - c \Delta u = 0 \) and so the above gives three wave equations, one for each component.

### 23.4.7 A Negative Observation

Many of the above applications of the divergence theorem are based on the assumption that matter is continuously distributed in a way that the above arguments are correct. In other words, a continuum. However, there is no such thing as a continuum. It has been known for some time now that matter is composed of atoms. It is not continuously distributed through some region of space as it is in the above. Apologists for this contradiction with reality sometimes say to consider enough of the material in question that it is reasonable to think of it as a continuum. This mystical reasoning is then violated as soon as they go from the integral form of the balance laws to the differential equations expressing the traditional formulation of these laws. See Problem 9 below, for example. However, these laws continue to be used and seem to lead to useful physical models which have value in predicting the behavior of physical systems. This is what justifies their use, not any fundamental truth.

### 23.4.8 Volumes Of Balls In \( \mathbb{R}^n \) (For Those Who Know About The Gamma Function)

Recall, \( B(x, r) \) denotes the set of all \( y \in \mathbb{R}^n \) such that \( |y - x| < r \). By the change of variables formula for multiple integrals or simple geometric reasoning, all balls of radius \( r \) have the same volume. Furthermore, simple reasoning or change of variables formula will show that the volume of the ball of radius \( r \) equals \( \alpha_n r^n \) where \( \alpha_n \) will denote the volume of
the unit ball in \( \mathbb{R}^n \). With the divergence theorem, it is now easy to give a simple relationship between the surface area of the ball of radius \( r \) and the volume. By the divergence theorem,
\[
\int_{B(0,r)} \text{div} \mathbf{x} \, dx = \int_{\partial B(0,r)} \mathbf{x} \cdot \mathbf{n} \, dA
\]
because the unit outward normal on \( \partial B(0,r) \) is \( \frac{x}{|x|} \). Therefore,
\[
nr^n = rA(\partial B(0,r))
\]
and so
\[
A(\partial B(0,r)) = nr^{n-1}.
\]
You recall the surface area of \( S^2 \equiv \{ x \in \mathbb{R}^3 \colon |x| = 1 \} \) is given by \( 4\pi r^2 \) while the volume of the ball, \( B(0,r) \) is \( \frac{4}{3}\pi r^3 \). This follows the above pattern. You just take the derivative with respect to the radius of the volume of the ball of radius \( r \) to get the area of the surface of this ball. Let \( \omega_n \) denote the area of the sphere \( S^{n-1} = \{ x \in \mathbb{R}^n \colon |x| = 1 \} \). I just showed that
\[
\omega_n = n\alpha_n.
\]

I want to find \( \alpha_n \) now and also to get a relationship between \( \omega_n \) and \( \omega_{n-1} \). Consider the following picture of the ball of radius \( \rho \) seen on the side.

![Diagram of a ball with slices](image)

Taking slices at height \( y \) as shown and using that these slices have \( n-1 \) dimensional area equal to \( \alpha_{n-1}r^{n-1} \), it follows
\[
\alpha_n\rho^n = 2\int_0^\rho \alpha_{n-1}(\rho^2 - y^2)^{(n-1)/2} \, dy
\]
In the integral, change variables, letting \( y = \rho \cos \theta \). Then
\[
\alpha_n\rho^n = 2\rho^n\alpha_{n-1}\int_0^{\pi/2} \sin^n(\theta) \, d\theta.
\]
It follows that
\[
\alpha_n = 2\alpha_{n-1}\int_0^{\pi/2} \sin^n(\theta) \, d\theta.
\]
Consequently,
\[
\omega_n = 2n\omega_{n-1}\int_0^{\pi/2} \sin^n(\theta) \, d\theta - (n-1)\int_0^{\pi/2} \sin^{n-2}(\theta) \, d\theta
\]
This is a little messier than I would like.
\[
\int_0^{\pi/2} \sin^n(\theta) \, d\theta = -\cos \theta \sin^{n-1} \theta|_0^{\pi/2} + (n-1)\int_0^{\pi/2} \cos^2 \theta \sin^{n-2} \theta \, d\theta
\]
\[
= (n-1)\int_0^{\pi/2} (1 - \sin^2 \theta) \sin^{n-2} (\theta) \, d\theta
\]
\[
= (n-1)\int_0^{\pi/2} \sin^{n-2} (\theta) \, d\theta - (n-1)\int_0^{\pi/2} \sin^n (\theta) \, d\theta
\]
Hence
\[ n \int_{0}^{\pi/2} \sin^n(\theta) \, d\theta = (n - 1) \int_{0}^{\pi/2} \sin^{n-2}(\theta) \, d\theta \quad (23.17) \]
and so (23.16) is of the form
\[ \omega_n = 2\omega_{n-1} \int_{0}^{\pi/2} \sin^{n-2}(\theta) \, d\theta. \quad (23.18) \]

So what is \( \alpha_n \) explicitly? Clearly \( \alpha_1 = 2 \) and \( \alpha_2 = \pi \).

**Theorem 23.4.2** \( \alpha_n = \frac{\pi^{n/2}}{\Gamma(n/2 + 1)} \) where \( \Gamma \) denotes the gamma function, defined for \( \alpha > 0 \) by
\[ \Gamma(\alpha) \equiv \int_{0}^{\infty} e^{-t} t^{\alpha-1} \, dt. \]

**Proof:** Recall that \( \Gamma(\alpha + 1) = \alpha \Gamma(\alpha) \). Now note the given formula holds if \( n = 1 \) because
\[ \Gamma \left( \frac{1}{2} + 1 \right) = \frac{1}{2} \Gamma \left( \frac{1}{2} \right) = \sqrt{\pi}. \]

(I leave it as an exercise for you to verify that \( \Gamma \left( \frac{1}{2} \right) = \sqrt{\pi} \).) Thus
\[ \alpha_1 = 2 = \frac{\sqrt{\pi}}{\sqrt{\pi/2}} \]
satisfying the formula. Now suppose this formula holds for \( k \leq n \). Then from the induction hypothesis, (23.18), (23.17), (23.15) and (23.16),
\[
\alpha_{n+1} = 2\alpha_n \int_{0}^{\pi/2} \sin^{n+1}(\theta) \, d\theta
\]
\[ = 2\alpha_n \frac{n}{n+1} \int_{0}^{\pi/2} \sin^{n-1}(\theta) \, d\theta
\]
\[ = 2\alpha_n \frac{n}{n+1} \frac{\alpha_{n-1}}{2\alpha_{n-2}}
\]
\[ = \frac{\pi^{n/2}}{\Gamma \left( \frac{n}{2} + 1 \right)} \frac{n}{n+1} \frac{1}{\Gamma \left( \frac{n-2}{2} + 1 \right)}
\]
\[ = \frac{\pi^{n/2}}{\Gamma \left( \frac{n-2}{2} + 1 \right)} \frac{n}{n+1} \frac{\Gamma \left( \frac{n-2}{2} + 1 \right)}{\Gamma \left( \frac{n-1}{2} + 1 \right)}
\]
\[ = 2\pi^{(n+1)/2} \frac{1}{n+1} \Gamma \left( \frac{n-1}{2} + 1 \right)
\]
\[ = \pi^{(n+1)/2} \frac{1}{\Gamma \left( \frac{n+1}{2} \right)} \frac{1}{\Gamma \left( \frac{n+1}{2} + 1 \right)}
\]
\[ = \pi^{(n+1)/2} \frac{\Gamma \left( \frac{n+1}{2} \right)}{\Gamma \left( \frac{n+1}{2} + 1 \right)}. \]

This proves the theorem.
23.4. SOME APPLICATIONS OF THE DIVERGENCE THEOREM

23.4.9 Electrostatics

Coloumb’s law says that the electric field intensity at \( x \) of a charge \( q \) located at point, \( x_0 \) is given by

\[
E = k \frac{q(x - x_0)}{|x - x_0|^3}
\]

where the electric field intensity is defined to be the force experienced by a unit positive charge placed at the point, \( x \). Note that this is a vector and that its direction depends on the sign of \( q \). It points away from \( x_0 \) if \( q \) is positive and points toward \( x_0 \) if \( q \) is negative. The constant, \( k \) is a physical constant like the gravitation constant. It has been computed through careful experiments similar to those used with the calculation of the gravitation constant.

The interesting thing about Coloumb’s law is that \( E \) is the gradient of a function. In fact,

\[
E = \nabla \left( \frac{qk}{|x - x_0|} \right).
\]

The other thing which is significant about this is that in three dimensions and for \( x \neq x_0 \),

\[
\nabla \cdot \nabla \left( \frac{qk}{|x - x_0|} \right) = \nabla \cdot E = 0. \tag{23.19}
\]

This is left as an exercise for you to verify.

These observations will be used to derive a very important formula for the integral,

\[
\int_{\partial U} E \cdot n \, dS
\]

where \( E \) is the electric field intensity due to a charge, \( q \) located at the point, \( x_0 \in U \), a bounded open set for which the divergence theorem holds.

Let \( U_\varepsilon \) denote the open set obtained by removing the open ball centered at \( x_0 \) which has radius \( \varepsilon \) where \( \varepsilon \) is small enough that the following picture is a correct representation of the situation.

Then on the boundary of \( B_\varepsilon \) the unit outer normal to \( U_\varepsilon \) is \( \frac{x - x_0}{|x - x_0|} \). Therefore,

\[
\int_{\partial B_\varepsilon} E \cdot n \, dS = - \int_{\partial B_\varepsilon} \frac{kq(x - x_0)}{|x - x_0|^3} \cdot \frac{x - x_0}{|x - x_0|^2} \, dS
\]

\[
= -kq \int_{\partial B_\varepsilon} \frac{1}{|x - x_0|^2} \, dS = -kq \int_{\partial B_\varepsilon} \frac{1}{\varepsilon^2} \, dS
\]

\[
= -kq \frac{4\pi \varepsilon^2}{\varepsilon^2} = -4\pi kq.
\]
Therefore, from the divergence theorem and observation (23.19),
\[-4\pi kq + \int_{\partial U} \mathbf{E} \cdot \mathbf{n} dS = \int_{\partial U_x} \mathbf{E} \cdot \mathbf{n} dS = \int_{U_x} \nabla \cdot \mathbf{E} dV = 0.\]

It follows that
\[4\pi kq = \int_{\partial U} \mathbf{E} \cdot \mathbf{n} dS.\]

If there are several charges located inside \( U \), say \( q_1, q_2, \cdots, q_n \), then letting \( E_i \) denote the electric field intensity of the \( i^{th} \) charge and \( \mathbf{E} \) denoting the total resulting electric field intensity due to all these charges,
\[\int_{\partial U} \mathbf{E} \cdot \mathbf{n} dS = \sum_{i=1}^{n} \int_{\partial U} E_i \cdot \mathbf{n} dS = \sum_{i=1}^{n} 4\pi kq_i = 4\pi k \sum_{i=1}^{n} q_i.\]

This is known as Gauss’s law and it is the fundamental result in electrostatics.

### 23.5 Exercises

1. To prove the divergence theorem, it was shown first that the spacial partial derivative in the volume integral could be exchanged for multiplication by an appropriate component of the exterior normal. This problem starts with the divergence theorem and goes the other direction. Assuming the divergence theorem holds for a region, \( V \), show that
\[\int_{\partial V} \mathbf{u} \cdot \mathbf{n} dA = \int_V \nabla \cdot \mathbf{u} dV.\]

2. Let \( V \) be such that the divergence theorem holds. Show that \( \int_V \nabla \cdot (u \nabla v) dV = \int_{\partial V} u \frac{\partial v}{\partial n} dA \) where \( \mathbf{n} \) is the exterior normal and \( \frac{\partial v}{\partial n} \) denotes the directional derivative of \( v \) in the direction \( \mathbf{n} \).

3. Let \( V \) be such that the divergence theorem holds. Show that \( \int \int_V (v \nabla^2 u - u \nabla^2 v) dV = \int \int_{\partial V} (v \frac{\partial u}{\partial n} - u \frac{\partial v}{\partial n}) dA \) where \( \mathbf{n} \) is the exterior normal and \( \frac{\partial v}{\partial n} \) is defined in Problem 2.

4. Let \( V \) be a ball and suppose \( \nabla^2 u = f \) in \( V \) while \( u = g \) on \( \partial V \). Show there is at most one solution to this boundary value problem which is \( C^2 \) in \( V \) and continuous on \( V \) with its boundary. **Hint:** You might consider \( w = u - v \) where \( u \) and \( v \) are solutions to the problem. Then use the result of Problem 2 and the identity
\[w \nabla^2 w = \nabla \cdot (w \nabla w) - \nabla w \cdot \nabla w\]
to conclude \( \nabla w = 0 \). Then show this implies \( w \) must be a constant by considering \( h(t) = w(tx + (1-t)y) \) and showing \( h \) is a constant. Alternatively, you might consider the maximum principle.

5. Show that \( \int_{\partial V} \nabla \times \mathbf{v} \cdot \mathbf{n} dA = 0 \) where \( V \) is a region for which the divergence theorem holds and \( \mathbf{v} \) is a \( C^2 \) vector field.

6. Let \( \mathbf{F}(x, y, z) = (x, y, z) \) be a vector field in \( \mathbb{R}^3 \) and let \( V \) be a three dimensional shape and let \( \mathbf{n} = (n_1, n_2, n_3) \). Show \( \int_{\partial V} (xn_1 + yn_2 + zn_3) dA = 3 \times \text{volume of } V \).

7. Does the divergence theorem hold for higher dimensions? If so, explain why it does. How about two dimensions?
8. Let \( \mathbf{F} = xi + yj + zk \) and let \( V \) denote the tetrahedron formed by the planes, \( x = 0, y = 0, z = 0, \) and \( \frac{1}{2}x + \frac{1}{2}y + \frac{1}{2}z = 1. \) Verify the divergence theorem for this example.

9. Suppose \( f : U \to \mathbb{R} \) is continuous where \( U \) is some open set and for all \( B \subseteq U \) where \( B \) is a ball, \( \int_B f(x) \, dV = 0. \) Show this implies \( f(x) = 0 \) for all \( x \in U. \)

10. Let \( U \) denote the box centered at \((0,0,0)\) with sides parallel to the coordinate planes which has width 4, length 2 and height 3. Find the flux integral \( \int_{\partial U} \mathbf{F} \cdot \mathbf{n} \, dS \) where \( \mathbf{F} = (x + 3, 2y, 3z). \) **Hint:** If you like, you might want to use the divergence theorem.

11. Verify (23.14) from (23.13) and the assumption that \( S = kF. \)

12. Fick’s law for diffusion states the flux of a diffusing species, \( J \) is proportional to the gradient of the concentration, \( c. \) Write this law getting the sign right for the constant of proportionality and derive an equation similar to the heat equation for the concentration, \( c. \) Typically, \( c \) is the concentration of some sort of pollutant or a chemical.

13. Show that if \( u_k, k = 1, 2, \cdots, n \) each satisfies (23.7) then for any choice of constants, \( c_1, \cdots, c_n, \) so does \( \sum_{k=1}^{n} c_k u_k. \)

14. Suppose \( k(x) = k, \) a constant and \( f = 0. \) Then in one dimension, the heat equation is of the form \( u_t = \alpha u_{xx}. \) Show \( u(x,t) = e^{-\alpha n^2 t} \sin(nx) \) satisfies the heat equation\(^2.\)

15. In a linear, viscous, incompressible fluid, the Cauchy stress is of the form

\[
T_{ij}(t, y) = \lambda \left( \frac{\partial v_i}{\partial y_j} + \frac{\partial v_j}{\partial y_i} \right)
\]

where the comma followed by an index indicates the partial derivative with respect to that variable and \( \mathbf{v} \) is the velocity. Thus

\[
v_{i,j} = \frac{\partial v_i}{\partial y_j}
\]

Show, using the balance of mass equation that incompressible implies \( \text{div} \mathbf{v} = 0. \) Next show the balance of momentum equation requires

\[
\rho \ddot{\mathbf{v}} - \frac{\lambda}{2} \Delta \mathbf{v} = \rho \left[ \frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial y_i} v_i \right] - \frac{\lambda}{2} \Delta \mathbf{v} = \mathbf{b}.
\]

This is the famous Navier Stokes equation for incompressible viscous linear fluids. There are still open questions related to this equation, one of which is worth \$1,000,000 at this time.

\(^2\)Fourier, an officer in Napoleon’s army studied solutions to the heat equation back in 1813. He was interested in heat flow in cannons. He sought to find solutions by adding up infinitely many solutions of this form. Actually, it was a little more complicated because cannons are not one dimensional but it was the beginning of the study of Fourier series, a topic which fascinated mathematicians for the next 150 years and motivated the development of analysis.
Stokes And Green’s Theorems

24.0.1 Outcomes

1. Recall and verify Green’s theorem.

2. Apply Green’s theorem to evaluate line integrals.

3. Apply Green’s theorem to find the area of a region.

4. Explain what is meant by the curl of a vector field.

5. Evaluate the curl of a vector field.

6. Derive and apply formulas involving divergence, gradient and curl.

7. Recall and use Stoke’s theorem.

8. Apply Stoke’s theorem to calculate the circulation or work of a vector field around a simple closed curve.

9. Recall and apply the fundamental theorem for line integrals.

10. Determine whether a vector field is a gradient using the curl test.

11. Recover a function from its gradient when possible.

24.1 Green’s Theorem

Green’s theorem is an important theorem which relates line integrals to integrals over a surface in the plane. It can be used to establish the much more significant Stoke’s theorem but is interesting for it’s own sake. Historically, it was important in the development of complex analysis. I will first establish Green’s theorem for regions of a particular sort and then show that the theorem holds for many other regions also. Suppose a region is of the form indicated in the following picture in which

\[
U = \{(x,y) : x \in (a, b) \text{ and } y \in (b(x), t(x))\}
\]

\[
= \{(x,y) : y \in (c, d) \text{ and } x \in (l(y), r(y))\}.
\]
I will refer to such a region as being convex in both the $x$ and $y$ directions.

**Lemma 24.1.1** Let $\mathbf{F}(x,y) \equiv (P(x,y), Q(x,y))$ be a $C^1$ vector field defined near $U$ where $U$ is a region of the sort indicated in the above picture which is convex in both the $x$ and $y$ directions. Suppose also that the functions, $r, l, t,$ and $b$ in the above picture are all $C^1$ functions and denote by $\partial U$ the boundary of $U$ oriented such that the direction of motion is counter clockwise. (As you walk around $U$ on $\partial U$, the points of $U$ are on your left.) Then

$$\int_{\partial U} P \, dx + Q \, dy \equiv$$

$$\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} = \int \int_{U} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA. \quad (24.1)$$

**Proof:** First consider the right side of (24.1).

$$\int \int_{U} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA =$$

$$= \int_{c}^{d} \int_{l(y)}^{r(y)} \frac{\partial Q}{\partial x} \, dx \, dy - \int_{a}^{b} \int_{r(x)}^{l(x)} \frac{\partial P}{\partial y} \, dy \, dx$$

$$= \int_{c}^{d} \left( Q(r(y), y) - Q(l(y), y) \right) dy + \int_{a}^{b} \left( P(x, b(x)) - P(x, t(x)) \right) dx. \quad (24.2)$$

Now consider the left side of (24.1). Denote by $V$ the vertical parts of $\partial U$ and by $H$ the horizontal parts.

$$\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} =$$

$$= \int_{\partial U} (\langle 0, Q \rangle + (P, 0)) \cdot d\mathbf{R}.$$

$$= \int_{c}^{d} \langle 0, Q(r(s), s) \rangle \cdot \langle r'(s), 1 \rangle ds + \int_{H} \langle 0, Q(r(s), s) \rangle \cdot \langle \pm 1, 0 \rangle ds$$

$$- \int_{c}^{d} \langle 0, Q(l(s), s) \rangle \cdot \langle l'(s), 1 \rangle ds + \int_{a}^{b} \langle P(s, b(s)), 0 \rangle \cdot \langle 1, b'(s) \rangle ds$$

$$+ \int_{V} \langle P(s, b(s)), 0 \rangle \cdot \langle 0, \pm 1 \rangle ds - \int_{a}^{b} \langle P(s, t(s)), 0 \rangle \cdot \langle 1, t'(s) \rangle ds$$

$$= \int_{c}^{d} Q(r(s), s) ds - \int_{c}^{d} Q(l(s), s) ds + \int_{a}^{b} P(s, b(s)) ds - \int_{a}^{b} P(s, t(s)) ds$$

which coincides with (24.2). This proves the lemma.
Corollary 24.1.2 Let everything be the same as in Lemma 24.1.1 but only assume the functions $r, l, t,$ and $b$ are continuous and piecewise $C^1$ functions. Then the conclusion this lemma is still valid.

Proof: The details are left for you. All you have to do is to break up the various line integrals into the sum of integrals over sub intervals on which the function of interest is $C^1$.

From this corollary, it follows (24.1) is valid for any triangle for example.

Now suppose (24.1) holds for $U_1, U_2, \ldots, U_m$ and the open sets, $U_k$ have the property that no two have nonempty intersection and their boundaries intersect only in a finite number of piecewise smooth curves. Then (24.1) must hold for $U \equiv \cup_{i=1}^m U_i$, the union of these sets. This is because

$$\int \int_U \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA =$$

$$= \sum_{k=1}^m \int \int_{U_k} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA$$

$$= \sum_{k=1}^m \int_{\partial U_k} \mathbf{F} \cdot d\mathbf{R} = \int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$$

because if $\Gamma = \partial U_k \cap \partial U_j$, then its orientation as a part of $\partial U_k$ is opposite to its orientation as a part of $\partial U_j$ and consequently the line integrals over $\Gamma$ will cancel, points of $\Gamma$ also not being in $\partial U$. As an illustration, consider the following picture for two such $U_k$.

Similarly, if $U \subseteq V$ and if also $\partial U \subseteq V$ and both $U$ and $V$ are open sets for which (24.1) holds, then the open set, $V \setminus (U \cup \partial U)$ consisting of what is left in $V$ after deleting $U$ along with its boundary also satisfies (24.1). Roughly speaking, you can drill holes in a region for which (24.1) holds and get another region for which this continues to hold provided (24.1) holds for the holes. To see why this is so, consider the following picture which typifies the situation just described.

Then

$$\int_{\partial V} \mathbf{F} \cdot d\mathbf{R} = \int \int_V \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA$$
\[
\begin{align*}
&= \int \int_U \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA + \int \int_{V \setminus U} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA \\
&= \int_{\partial U} F \cdot dR + \int \int_{V \setminus U} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA
\end{align*}
\]

and so
\[
\int \int_{V \setminus U} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \int_{\partial V} F \cdot dR - \int_{\partial U} F \cdot dR
\]

which equals
\[
\int_{\partial (V \setminus U)} F \cdot dR
\]

where \( \partial V \) is oriented as shown in the picture. (If you walk around the region, \( V \setminus U \) with the area on the left, you get the indicated orientation for this curve.)

You can see that (24.1) is valid quite generally. This verifies the following theorem.

**Theorem 24.1.3** (Green’s Theorem) Let \( U \) be an open set in the plane and let \( \partial U \) be piecewise smooth and let \( \langle P(x, y), Q(x, y) \rangle \) be a \( C^1 \) vector field defined near \( U \). Then it is often the case that

\[
\int_{\partial U} F \cdot dR = \int \int_U \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) (x, y) \ dA.
\]

**Proposition 24.1.4** Let \( U \) be an open set in \( \mathbb{R}^2 \) for which Green’s theorem holds. Then

\[
\text{Area of } U = \int_{\partial U} F \cdot dR
\]

where \( F(x, y) = \frac{1}{2} ( -y, x ), (0, x), \text{ or } ( -y, 0) \).

**Proof:** This follows immediately from Green’s theorem.

**Example 24.1.5** Use Proposition 24.1.4 to find the area of the ellipse

\[
\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1.
\]

You can parameterize the boundary of this ellipse as

\[
x = a \cos t, \ y = b \sin t, \ t \in [0, 2\pi].
\]

Then from Use Proposition 24.1.4,

\[
\text{Area equals} = \frac{1}{2} \int_0^{2\pi} ( -b \sin t, a \cos t ) \cdot ( -a \sin t, b \cos t ) \ dt
\]

\[
= \frac{1}{2} \int_0^{2\pi} (ab) \ dt = \pi ab.
\]

**Example 24.1.6** Find \( \int_{\partial U} F \cdot dR \) where \( U \) is the set, \( \{(x, y) : x^2 + 3y^2 \leq 9\} \) and \( F(x, y) = (y, -x) \).

\footnote{For a general version see the advanced calculus book by Apostol. The general versions involve the concept of a rectifiable Jordan curve.}
One way to do this is to parameterize the boundary of $U$ and then compute the line integral directly. It is easier to use Green’s theorem. The desired line integral equals

$$\int \int_U ((-1) - 1) \, dA = -2 \int \int_U \, dA.$$ 

Now $U$ is an ellipse having area equal to $3\sqrt{3}$ and so the answer is $-6\sqrt{3}$.

**Example 24.1.7** Find $\int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$ where $U$ is the set, $\{(x, y) : 2 \leq x \leq 4, 0 \leq y \leq 3\}$ and $\mathbf{F}(x, y) = (x \sin y, y^3 \cos x)$.

From Green’s theorem this line integral equals

$$\int_2^4 \int_0^3 (-y^3 \sin x - x \cos y) \, dy \, dx = \frac{81}{4} \cos 4 - 6 \sin 3 - \frac{81}{4} \cos 2.$$ 

This is much easier than computing the line integral because you don’t have to break the boundary in pieces and consider each separately.

**Example 24.1.8** Find $\int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$ where $U$ is the set, $\{(x, y) : 2 \leq x \leq 4, x \leq y \leq 3\}$ and $\mathbf{F}(x, y) = (x \sin y, y \sin x)$.

From Green’s theorem this line integral equals

$$\int_2^4 \int_x^3 (y \cos x - x \cos y) \, dy \, dx = -\frac{3}{2} \sin 4 - 6 \sin 3 - 8 \cos 4 - \frac{9}{2} \sin 2 + 4 \cos 2.$$ 

### 24.2 Stoke’s Theorem From Green’s Theorem

Stoke’s theorem is a generalization of Green’s theorem which relates the integral over a surface to the integral around the boundary of the surface. These terms are a little different from what occurs in $\mathbb{R}^2$. To describe this, consider a sock. The surface is the sock and its boundary will be the edge of the opening of the sock in which you place your foot. Another way to think of this is to imagine a region in $\mathbb{R}^2$ of the sort discussed above for Green’s theorem. Suppose it is on a sheet of rubber and the sheet of rubber is stretched in three dimensions. The boundary of the resulting surface is the result of the stretching applied to the boundary of the original region in $\mathbb{R}^2$. Here is a picture describing the situation.

Recall the following definition of the curl of a vector field.
Definition 24.2.1 Let
\[ F(x, y, z) = (F_1(x, y, z), F_2(x, y, z), F_3(x, y, z)) \]
be a \( C^1 \) vector field defined on an open set, \( V \) in \( \mathbb{R}^3 \). Then
\[
\nabla \times F \equiv \begin{vmatrix}
  \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\
  F_1 & F_2 & F_3
\end{vmatrix}
\]
\[
= \left( \frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} \right) i + \left( \frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x} \right) j + \left( \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) k.
\]
This is also called \( \text{curl} (F) \) and written as indicated, \( \nabla \times F \).

The following lemma gives the fundamental identity which will be used in the proof of Stoke’s theorem.

Lemma 24.2.2 Let \( R : U \to V \subseteq \mathbb{R}^3 \) where \( U \) is an open subset of \( \mathbb{R}^2 \) and \( V \) is an open subset of \( \mathbb{R}^3 \). Suppose \( R \) is \( C^2 \) and let \( F \) be a \( C^1 \) vector field defined in \( V \).

\[
(\mathbf{R}_u \times \mathbf{R}_v) \cdot (\nabla \times F) (\mathbf{R}(u, v)) = (\mathbf{F} \circ \mathbf{R})_u \cdot \mathbf{R}_v - (\mathbf{F} \circ \mathbf{R})_v \cdot \mathbf{R}_u) (u, v). \tag{24.3}
\]

Proof: Start with the left side and let \( x_i = R_i(u, v) \) for short.

\[
(\mathbf{R}_u \times \mathbf{R}_v) \cdot (\nabla \times F) (\mathbf{R}(u, v)) = \epsilon_{ijk} x_{ju} x_{kv} \partial F_s \partial x_r
\]
\[
= (\delta_{jr} \delta_{ks} - \delta_{js} \delta_{kr}) x_{ju} x_{kv} \partial F_s \partial x_r
\]
\[
= x_{ju} x_{kv} \partial F_k \partial x_j - x_{ju} x_{kv} \partial F_j \partial x_k
\]
\[
= \mathbf{R}_u \cdot \frac{\partial (\mathbf{F} \circ \mathbf{R})}{\partial u} - \mathbf{R}_v \cdot \frac{\partial (\mathbf{F} \circ \mathbf{R})}{\partial v}
\]
which proves (24.3).

The proof of Stoke’s theorem given next follows [7]. First, it is convenient to give a definition.

Definition 24.2.3 A vector valued function, \( \mathbf{R} : U \subseteq \mathbb{R}^m \to \mathbb{R}^n \) is said to be in \( C^k (U, \mathbb{R}^n) \) if it is the restriction to \( U \) of a vector valued function which is defined on \( \mathbb{R}^m \) and is \( C^k \). That is this function has continuous partial derivatives up to order \( k \).

Theorem 24.2.4 (Stoke’s Theorem) Let \( U \) be any region in \( \mathbb{R}^2 \) for which the conclusion of Green’s theorem holds and let \( R \in C^2 (U, \mathbb{R}^3) \) be a one to one function satisfying \( |(\mathbf{R}_u \times \mathbf{R}_v) (u, v)| \neq 0 \) for all \( (u, v) \in U \) and let \( S \) denote the surface,

\[
\begin{align*}
S & \equiv \{ \mathbf{R} (u, v) : (u, v) \in U \} , \\
\partial S & \equiv \{ \mathbf{R} (u, v) : (u, v) \in \partial U \}
\end{align*}
\]

where the orientation on \( \partial S \) is consistent with the counter clockwise orientation on \( \partial U \) (\( U \) is on the left as you walk around \( \partial U \)). Then for \( \mathbf{F} \) a \( C^1 \) vector field defined near \( S \),

\[
\int_{\partial S} \mathbf{F} \cdot d\mathbf{R} = \int_S \text{curl} (\mathbf{F}) \cdot n dS
\]

where \( n \) is the normal to \( S \) defined by

\[
n \equiv \frac{\mathbf{R}_u \times \mathbf{R}_v}{|\mathbf{R}_u \times \mathbf{R}_v|}
\]
24.2. STOKE’S THEOREM FROM GREEN’S THEOREM

Proof: Letting \( C \) be an oriented part of \( \partial U \) having parametrization, \( r(t) \equiv (u(t), v(t)) \) for \( t \in [\alpha, \beta] \) and letting \( R(C) \) denote the oriented part of \( \partial S \) corresponding to \( C \),

\[
\int_{R(C)} F \cdot dR = \int_{\alpha}^{\beta} F(R(u(t), v(t))) \cdot (R_u u'(t) + R_v v'(t)) \, dt
\]

\[
= \int_{\alpha}^{\beta} F(R(u(t), v(t))) R_u (u(t), v(t)) \, u'(t) \, dt + \int_{\alpha}^{\beta} F(R(u(t), v(t))) R_v (u(t), v(t)) \, v'(t) \, dt
\]

\[
= \int_{C} \langle (F \circ R) \cdot R_u, (F \circ R) \cdot R_v \rangle \cdot dr.
\]

Since this holds for each such piece of \( \partial U \), it follows

\[
\int_{\partial S} F \cdot dR = \int_{\partial U} \langle (F \circ R) \cdot R_u, (F \circ R) \cdot R_v \rangle \cdot dr.
\]

By the assumption that the conclusion of Green’s theorem holds for \( U \), this equals

\[
\int \int_{U} \left[ ((F \circ R) \cdot R_v)_u - ((F \circ R) \cdot R_u)_v \right] \, dA
\]

\[
= \int \int_{U} \left[ (F \circ R)_u \cdot R_v + (F \circ R) \cdot R_{uv} - (F \circ R)_v \cdot R_u \right] \, dA
\]

\[
= \int \int_{U} \left[ (F \circ R)_u \cdot R_v - (F \circ R)_v \cdot R_u \right] \, dA
\]

the last step holding by equality of mixed partial derivatives, a result of the assumption that \( R \) is \( C^2 \). Now by Lemma 24.2.2, this equals

\[
\int \int_{U} (R_u \times R_v) \cdot (\nabla \times F) \, dA
\]

\[
= \int \int_{U} \nabla \times F \cdot (R_u \times R_v) \, dA
\]

\[
= \int \int_{S} \nabla \times F \cdot n \, dS
\]

because \( dS = |(R_u \times R_v)| \, dA \) and \( n = \frac{(R_u \times R_v)}{|R_u \times R_v|} \). Thus

\[
(R_u \times R_v) \, dA = \frac{(R_u \times R_v)}{|R_u \times R_v|} \, |(R_u \times R_v)| \, dA
\]

\[
= ndS.
\]

This proves Stoke’s theorem.

Note that there is no mention made in the final result that \( R \) is \( C^2 \). Therefore, it is not surprising that versions of this theorem are valid in which this assumption is not present. It is possible to obtain extremely general versions of Stoke’s theorem if you use the Lebesgue integral.
24.2.1 Orientation

It turns out there are more general formulations of Stoke’s theorem than what is presented above. However, it is always necessary for the surface, $S$, to be orientable. This means it is possible to obtain a vector field for a unit normal to the surface which is a continuous function of position on $S$. An example of a surface which is not orientable is the famous Möbius band, obtained by taking a long rectangular piece of paper and gluing the ends together after putting a twist in it. Here is a picture of one.

There is something quite interesting about this Möbius band and this is that it can be written parametrically with a simple parameter domain. The picture above is a Maple graph of the parametrically defined surface

$$R(\theta, v) \equiv \begin{cases} x = 4 \cos \theta + v \cos \frac{\theta}{2}, \\ y = 4 \sin \theta + v \cos \frac{\theta}{2}, \\ z = v \sin \frac{\theta}{2}, \end{cases}, \theta \in [0, 2\pi], v \in [-1, 1].$$

An obvious question is why the normal vector, $R_{,\theta} \times R_{,v}/|R_{,\theta} \times R_{,v}|$ is not a continuous function of position on $S$. You can see easily that it is a continuous function of both $\theta$ and $v$. However, the map, $R$, is not one to one. In fact, $R(0, 0) = R(2\pi, 0)$. Therefore, near this point on $S$, there are two different values for the above normal vector. In fact, a short computation will show this normal vector is

$$\frac{(4 \sin \frac{1}{2} \theta \cos \theta - \frac{1}{2} v, 4 \sin \frac{1}{2} \theta \sin \theta + \frac{1}{2} v, -8 \cos^2 \frac{1}{2} \theta \sin \frac{1}{2} \theta - 8 \cos^3 \frac{1}{2} \theta + 4 \cos \frac{1}{2} \theta)}{\sqrt{16 \sin^2 \left(\frac{\theta}{2}\right) + v^2 + 4 \sin \left(\frac{\theta}{2}\right) v (\sin \theta - \cos \theta) + (-8 \cos^2 \frac{1}{2} \theta \sin \frac{1}{2} \theta - 8 \cos^3 \frac{1}{2} \theta + 4 \cos \frac{1}{2} \theta)^2}},$$

and you can verify that the denominator will not vanish. Letting $v = 0$ and $\theta = 0$ and $2\pi$ yields the two vectors,

$$(0, 0, -1), (0, 0, 1)$$

so there is a discontinuity. This is why I was careful to say in the statement of Stoke’s theorem given above that $R$ is one to one.

The Möbius band has some usefulness. In old machine shops the equipment was run by a belt which was given a twist to spread the surface wear on the belt over twice the area.

The above explanation shows that $R_{,\theta} \times R_{,v}/|R_{,\theta} \times R_{,v}|$ fails to deliver an orientation for the Möbius band. However, this does not answer the question whether there is some orientation for it other than this one. In fact there is none. You can see this by looking at the first of the two pictures below or by making one and tracing it with a pencil. There is only one side to the Möbius band. An oriented surface must have two sides, one side identified by the given unit normal which varies continuously over the surface and the other side identified by the negative of this normal. The second picture below was taken by Dr. Ouyang when he was at meetings in Paris and saw it at a museum.
24.2.2 Conservative Vector Fields

Definition 24.2.5 A vector field, $\mathbf{F}$ defined in a three dimensional region is said to be conservative if for every piecewise smooth closed curve, $C$, it follows $\int_C \mathbf{F} \cdot d\mathbf{R} = 0$.

Definition 24.2.6 Let $(\mathbf{x}, \mathbf{p}_1, \cdots, \mathbf{p}_n, \mathbf{y})$ be an ordered list of points in $\mathbb{R}^p$. Let

$$p(\mathbf{x}, \mathbf{p}_1, \cdots, \mathbf{p}_n, \mathbf{y})$$

denote the piecewise smooth curve consisting of a straight line segment from $\mathbf{x}$ to $\mathbf{p}_1$ and then the straight line segment from $\mathbf{p}_1$ to $\mathbf{p}_2$ ··· and finally the straight line segment from $\mathbf{p}_n$ to $\mathbf{y}$. This is called a polygonal curve. An open set in $\mathbb{R}^p$, $U$, is said to be a region if it has the property that for any two points, $\mathbf{x}, \mathbf{y} \in U$, there exists a polygonal curve joining the two points.

Conservative vector fields are important because of the following theorem, sometimes called the fundamental theorem for line integrals.

Theorem 24.2.7 Let $U$ be a region in $\mathbb{R}^p$ and let $\mathbf{F} : U \to \mathbb{R}^p$ be a continuous vector field. Then $\mathbf{F}$ is conservative if and only if there exists a scalar valued function of $p$ variables, $\phi$ such that $\mathbf{F} = \nabla \phi$. Furthermore, if $C$ is an oriented curve which goes from $\mathbf{x}$ to $\mathbf{y}$ in $U$, then

$$\int_C \mathbf{F} \cdot d\mathbf{R} = \phi(\mathbf{y}) - \phi(\mathbf{x}). \tag{24.4}$$

Thus the line integral is path independent in this case. This function, $\phi$ is called a scalar potential for $\mathbf{F}$.

Proof: To save space and fussing over things which are unimportant, denote by $p(\mathbf{x}_0, \mathbf{x})$ a polygonal curve from $\mathbf{x}_0$ to $\mathbf{x}$. Thus the orientation is such that it goes from $\mathbf{x}_0$ to $\mathbf{x}$. The curve $p(\mathbf{x}, \mathbf{x}_0)$ denotes the same set of points but in the opposite order. Suppose first $\mathbf{F}$ is conservative. Fix $\mathbf{x}_0 \in U$ and let

$$\phi(\mathbf{x}) \equiv \int_{p(\mathbf{x}_0, \mathbf{x})} \mathbf{F} \cdot d\mathbf{R}.$$
This is well defined because if \( q(x_0, x) \) is another polygonal curve joining \( x_0 \) to \( x \), then the curve obtained by following \( p(x_0, x) \) from \( x_0 \) to \( x \) and then from \( x \) to \( x_0 \) along \( q(x, x_0) \) is a closed piecewise smooth curve and so by assumption, the line integral along this closed curve equals 0. However, this integral is just

\[
\int_{p(x_0, x)} \mathbf{F} \cdot d\mathbf{R} + \int_{q(x, x_0)} \mathbf{F} \cdot d\mathbf{R} = \int_{p(x_0, x)} \mathbf{F} \cdot d\mathbf{R} - \int_{q(x, x_0)} \mathbf{F} \cdot d\mathbf{R}
\]

which shows

\[
\int_{p(x_0, x)} \mathbf{F} \cdot d\mathbf{R} = \int_{q(x, x_0)} \mathbf{F} \cdot d\mathbf{R}
\]

and that \( \phi \) is well defined. For small \( t \),

\[
\frac{\phi(x + t\mathbf{e}_i) - \phi(x)}{t} = \int_{p(x_0, x + t\mathbf{e}_i)} \mathbf{F} \cdot d\mathbf{R} - \int_{p(x_0, x)} \mathbf{F} \cdot d\mathbf{R}
\]

\[
= \int_{p(x_0, x)} \mathbf{F} \cdot d\mathbf{R} + \int_{p(x_0, x + t\mathbf{e}_i)} \mathbf{F} \cdot d\mathbf{R} - \int_{p(x_0, x)} \mathbf{F} \cdot d\mathbf{R}
\]

Since \( U \) is open, for small \( t \), the ball of radius \( |t| \) centered at \( x \) is contained in \( U \). Therefore, the line segment from \( x \) to \( x + t\mathbf{e}_i \) is also contained in \( U \) and so one can take \( p(x, x + t\mathbf{e}_i) = x + s(t\mathbf{e}_i) \) for \( s \in [0, 1] \). Therefore, the above difference quotient reduces to

\[
\frac{1}{t} \int_0^t \mathbf{F}(x + s(t\mathbf{e}_i)) \cdot t\mathbf{e}_i \, ds = \int_0^1 F_i(x + s(t\mathbf{e}_i)) \, ds
\]

by the mean value theorem for integrals. Here \( s_t \) is some number between 0 and 1. By continuity of \( \mathbf{F} \), this converges to \( F_i(x) \) as \( t \to 0 \). Therefore, \( \nabla \phi = \mathbf{F} \) as claimed.

Conversely, if \( \nabla \phi = \mathbf{F} \), then if \( \mathbf{R} : [a, b] \to \mathbb{R}^p \) is any \( C^1 \) curve joining \( x \) to \( y \),

\[
\int_a^b \mathbf{F}(\mathbf{R}(t)) \cdot \mathbf{R}'(t) \, dt = \int_a^b \nabla \phi(\mathbf{R}(t)) \cdot \mathbf{R}'(t) \, dt
\]

\[
= \int_a^b \frac{d}{dt}(\phi(\mathbf{R}(t))) \, dt
\]

\[
= \phi(\mathbf{R}(b)) - \phi(\mathbf{R}(a))
\]

\[
= \phi(y) - \phi(x)
\]

and this verifies (24.4) in the case where the curve joining the two points is smooth. The general case follows immediately from this by using this result on each of the pieces of the piecewise smooth curve. For example if the curve goes from \( x \) to \( p \) and then from \( p \) to \( y \), the above would imply the integral over the curve from \( x \) to \( p \) is \( \phi(p) - \phi(x) \) while from \( p \) to \( y \) the integral would yield \( \phi(y) - \phi(p) \). Adding these gives \( \phi(y) - \phi(x) \). The formula (24.4) implies the line integral over any closed curve equals zero because the starting and ending points of such a curve are the same. This proves the theorem.

**Example 24.2.8** Let \( \mathbf{F}(x, y, z) = (\cos x - yz \sin(xz), \cos(xz), -yz \sin(xz)) \). Let \( C \) be a piecewise smooth curve which goes from \((\pi, 1, 1)\) to \((\frac{\pi}{2}, 3, 2)\). Find \( \int_C \mathbf{F} \cdot d\mathbf{R} \).

The specifics of the curve are not given so the problem is nonsense unless the vector field is conservative. Therefore, it is reasonable to look for the function, \( \phi \) satisfying \( \nabla \phi = \mathbf{F} \). Such a function satisfies

\[
\phi_z = \cos x - y (\sin xz) z
\]
and so, assuming $\phi$ exists,

$$
\phi(x, y, z) = \sin x + y \cos (xz) + \psi(y, z).
$$

I have to add in the most general thing possible, $\psi(y, z)$ to ensure possible solutions are not being thrown out. It wouldn’t be good at this point to add in a constant since the answer could involve a function of either or both of the other variables. Now from what was just obtained,

$$
\phi_y = \cos (xz) + \psi_y = \cos xz
$$

and so it is possible to take $\psi_y = 0$. Consequently, $\phi$, if it exists is of the form

$$
\phi(x, y, z) = \sin x + y \cos (xz) + \psi(z).
$$

Now differentiating this with respect to $z$ gives

$$
\phi_z = -yx \sin (xz) + \psi_z = -yx \sin (xz)
$$

and this shows $\psi$ does not depend on $z$ either. Therefore, it suffices to take $\psi = 0$ and

$$
\phi(x, y, z) = \sin x + y \cos (xz).
$$

Therefore, the desired line integral equals

$$
\sin \left( \frac{\pi}{2} \right) + 3 \cos (\pi) - (\sin (\pi) + \cos (\pi)) = -1.
$$

The above process for finding $\phi$ will not lead you astray in the case where there does not exist a scalar potential. As an example, consider the following.

**Example 24.2.9** Let $\mathbf{F}(x, y, z) = (x, y^2, x)$. Find a scalar potential for $\mathbf{F}$ if it exists.

If $\phi$ exists, then $\phi_x = x$ and so $\phi = \frac{x^2}{2} + \psi(y, z)$. Then $\phi_y = \psi_y(y, z) = xy^2$ but this is impossible because the left side depends only on $y$ and $z$ while the right side depends also on $x$. Therefore, this vector field is not conservative and there does not exist a scalar potential.

**Definition 24.2.10** A set of points in three dimensional space, $V$ is simply connected if every piecewise smooth closed curve, $C$ is the edge of a surface, $S$ which is contained entirely within $V$ in such a way that Stokes theorem holds for the surface, $S$ and its edge, $C$. 

![Diagram](attachment:image.png)
This is like a sock. The surface is the sock and the curve, \( C \) goes around the opening of the sock.

As an application of Stoke’s theorem, here is a useful theorem which gives a way to check whether a vector field is conservative.

**Theorem 24.2.11** For a three dimensional simply connected open set, \( V \) and \( F \) a \( C^1 \) vector field defined in \( V \), \( F \) is conservative if \( \nabla \times F = 0 \) in \( V \).

**Proof:** If \( \nabla \times F = 0 \) then taking an arbitrary closed curve, \( C \), and letting \( S \) be a surface bounded by \( C \) which is contained in \( V \), Stoke’s theorem implies

\[
0 = \int_S \nabla \times F \cdot n \, dA = \int_C F \cdot dR.
\]

Thus \( F \) is conservative.

**Example 24.2.12** Determine whether the vector field,

\[
(4x^3 + 2(\cos(x^2 + z^2)) \, x) \text{, } 1 \text{, } 2(\cos(x^2 + z^2)) \, z)
\]

is conservative.

Since this vector field is defined on all of \( \mathbb{R}^3 \), it only remains to take its curl and see if it is the zero vector.

\[
\begin{vmatrix}
  \mathbf{i} & \mathbf{j} & \mathbf{k} \\
  \partial_x & \partial_y & \partial_z \\
  4x^3 + 2(\cos(x^2 + z^2)) \, x & 1 & 2(\cos(x^2 + z^2)) \, z
\end{vmatrix}.
\]

This is obviously equal to zero. Therefore, the given vector field is conservative. Can you find a potential function for it? Let \( \phi \) be the potential function. Then \( \phi_z = 2(\cos(x^2 + z^2)) \, z \) and so \( \phi(x, y, z) = \sin(x^2 + z^2) + g(x, y) \). Now taking the derivative of \( \phi \) with respect to \( y \), you see \( g_y = 1 \) so \( g(x, y) = y + h(x) \). Hence \( \phi(x, y, z) = y + g(x) + \sin(x^2 + z^2) \). Taking the derivative with respect to \( x \), you get \( 4x^3 + 2(\cos(x^2 + z^2)) \, x = g'(x) + 2x \cos(x^2 + z^2) \) and so it suffices to take \( g(x) = x^4 \). Hence \( \phi(x, y, z) = y + x^4 + \sin(x^2 + z^2) \).

**24.2.3 Some Terminology**

If \( F = (P, Q, R) \) is a vector field. Then the statement that \( F \) is conservative is the same as saying the differential form \( P \, dx + Q \, dy + R \, dz \) is exact. Some people like to say things in terms of vector fields and some say it in terms of differential forms. In Example 24.2.12, the differential form \( (4x^3 + 2(\cos(x^2 + z^2)) \, x) \, dx + dy + (2(\cos(x^2 + z^2)) \, z) \, dz \) is exact.

**24.2.4 Maxwell’s Equations And The Wave Equation**

Many of the ideas presented above are useful in analyzing Maxwell’s equations. These equations are derived in advanced physics courses. They are

\[
\nabla \times \mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{B}}{\partial t} = 0
\]

(24.5)

\[
\nabla \cdot \mathbf{E} = 4\pi \rho
\]

(24.6)

\[
\nabla \times \mathbf{B} - \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t} = 4\pi \mathbf{f}
\]

(24.7)

\[
\nabla \cdot \mathbf{B} = 0
\]

(24.8)
and it is assumed these hold on all of $\mathbb{R}^3$ to eliminate technical considerations having to do with whether something is simply connected.

In these equations, $\mathbf{E}$ is the electrostatic field and $\mathbf{B}$ is the magnetic field while $\rho$ and $\mathbf{f}$ are sources. By (24.8) $\mathbf{B}$ has a vector potential, $\mathbf{A}_1$ such that $\mathbf{B} = \nabla \times \mathbf{A}_1$. Now go to (24.5) and write
\[
\nabla \times \mathbf{E} + \frac{1}{c} \nabla \times \frac{\partial \mathbf{A}_1}{\partial t} = 0
\]
showing that
\[
\nabla \times \left( \mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{A}_1}{\partial t} \right) = 0
\]
It follows $\mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{A}_1}{\partial t}$ has a scalar potential, $\psi_1$ satisfying
\[
\nabla \psi_1 = \mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{A}_1}{\partial t}.
\] (24.9)

Now suppose $\phi$ is a time dependent scalar field satisfying
\[
\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = \frac{1}{c} \frac{\partial \psi_1}{\partial t} - \nabla \cdot \mathbf{A}_1.
\] (24.10)

Next define
\[
\mathbf{A} \equiv \mathbf{A}_1 + \nabla \phi, \quad \psi \equiv \psi_1 + \frac{1}{c} \frac{\partial \phi}{\partial t}.
\] (24.11)

Therefore, in terms of the new variables, (24.10) becomes
\[
\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = \frac{1}{c} \left( \frac{\partial \psi}{\partial t} - \frac{1}{c} \frac{\partial^2 \phi}{\partial t^2} \right) - \nabla \cdot \mathbf{A} + \nabla^2 \phi
\]
which yields
\[
0 = \frac{\partial \psi}{\partial t} - c \nabla \cdot \mathbf{A}.
\] (24.12)

Then it follows from Theorem 23.1.3 on Page 422 that $\mathbf{A}$ is also a vector potential for $\mathbf{B}$. That is
\[
\nabla \times \mathbf{A} = \mathbf{B}.
\] (24.13)

From (24.9)
\[
\nabla \left( \psi - \frac{1}{c} \frac{\partial \phi}{\partial t} \right) = \mathbf{E} + \frac{1}{c} \left( \frac{\partial \mathbf{A}}{\partial t} - \nabla \frac{\partial \phi}{\partial t} \right)
\]
and so
\[
\nabla \psi = \mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{A}}{\partial t}.
\] (24.14)

Using (24.7) and (24.14),
\[
\nabla \times (\nabla \times \mathbf{A}) - \frac{1}{c} \frac{\partial}{\partial t} \left( \nabla \psi - \frac{1}{c} \frac{\partial \mathbf{A}}{\partial t} \right) = \frac{4\pi}{c} \mathbf{f}.
\] (24.15)

Now from Theorem 23.1.3 on Page 422 this implies
\[
\nabla (\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} - \nabla \left( \frac{1}{c} \frac{\partial \psi}{\partial t} \right) + \frac{1}{c^2} \frac{\partial^2 \mathbf{A}}{\partial t^2} = \frac{4\pi}{c} \mathbf{f}
\]
and using (24.12), this gives
\[
\frac{1}{c^2} \frac{\partial^2 \mathbf{A}}{\partial t^2} - \nabla^2 \mathbf{A} = \frac{4\pi}{c} \mathbf{f}.
\] (24.16)
Also from (24.14), (24.6), and (24.12),

\[ \nabla^2 \psi = \nabla \cdot \mathbf{E} + \frac{1}{c} \frac{\partial}{\partial t} (\nabla \cdot \mathbf{A}) \]

\[ = 4\pi \rho + \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} \]

and so

\[ \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} - \nabla^2 \psi = -4\pi \rho. \quad (24.17) \]

This is very interesting. If a solution to the wave equations, (24.17), and (24.16) can be found along with a solution to (24.12), then letting the magnetic field be given by (24.13) and letting \( \mathbf{E} \) be given by (24.14) the result is a solution to Maxwell’s equations. This is significant because wave equations are easier to think of than Maxwell’s equations. Note the above argument also showed that it is always possible, by solving another wave equation, to get (24.12) to hold.

### 24.3 Exercises

1. Determine whether the vector field, \((2xy^3 \sin z^4, 3x^2y^2 \sin z^4 + 1, 4x^2y^3 (\cos z^4) z^3 + 1)\) is conservative. If it is conservative, find a potential function.

2. Determine whether the vector field, \((2xy^3 \sin z + y^2 + z, 3x^2y^2 \sin z + 2xy, x^2y^3 \cos z + x)\) is conservative. If it is conservative, find a potential function.

3. Determine whether the vector field, \((2xy^3 \sin z + z, 3x^2y^2 \sin z + 2xy, x^2y^3 \cos z + x)\) is conservative. If it is conservative, find a potential function.

4. Find scalar potentials for the following vector fields if it is possible to do so. If it is not possible to do so, explain why.

   (a) \((y^2, 2xy + \sin z, 2z + y \cos z)\)

   (b) \((2z (\cos (x^2 + y^2)) x, 2z (\cos (x^2 + y^2)) y, \sin (x^2 + y^2) + 2z)\)

   (c) \((f(x), g(y), h(z))\)

   (d) \((xy, z^2, y^3)\)

   (e) \((z + 2\frac{x^2}{x^2 + y^2 + 1, 2\frac{y^2}{x^2 + y^2 + 1}}, x + 3z^2)\)

5. If a vector field is not conservative on the set \(U\), is it possible the same vector field could be conservative on some subset of \(U\)? Explain and give examples if it is possible. If it is not possible also explain why.

6. Prove that if a vector field, \(\mathbf{F}\) has a scalar potential, then it has infinitely many scalar potentials.

7. Here is a vector field: \(\mathbf{F} \equiv (2xy, x^2 - 5y^4, 3z^2)\). Find \(\int_C \mathbf{F} \cdot d\mathbf{R}\) where \(C\) is a curve which goes from \((1, 2, 3)\) to \((4, -2, 1)\).

8. Here is a vector field: \(\mathbf{F} \equiv (2xy, x^2 - 5y^4, 3(\cos z^3) z^2)\). Find \(\int_C \mathbf{F} \cdot d\mathbf{R}\) where \(C\) is a curve which goes from \((1, 0, 1)\) to \((-4, -2, 1)\).

9. Find \(\int_{\partial U} \mathbf{F} \cdot d\mathbf{R}\) where \(U\) is the set, \(\{(x, y) : 2 \leq x \leq 4, 0 \leq y \leq x\}\) and \(\mathbf{F}(x, y) = (x \sin y, y \sin x)\).
10. Find $\int_{\partial U} \mathbf{F} \cdot d\mathbf{r}$ where $U$ is the set, \( \{(x, y) : 2 \leq x \leq 3, 0 \leq y \leq x^2\} \) and \( \mathbf{F}(x, y) = (x \cos y, y + x) \).

11. Find $\int_{\partial U} \mathbf{F} \cdot d\mathbf{r}$ where $U$ is the set, \( \{(x, y) : 1 \leq x \leq 2, x \leq y \leq 3\} \) and \( \mathbf{F}(x, y) = (x \sin y, y \sin x) \).

12. Find $\int_{\partial U} \mathbf{F} \cdot d\mathbf{r}$ where $U$ is the set, \( \{(x, y) : x^2 + y^2 \leq 2\} \) and \( \mathbf{F}(x, y) = (-y^3, x^3) \).

13. Show that for many open sets in $\mathbb{R}^2$, Area of $U = \int_{\partial U} y \, dx$, and Area of $U = \int_{\partial U} -y \, dx$ and Area of $U = \frac{1}{2} \int_{\partial U} -y \, dx + x \, dy$. Hint: Use Green’s theorem.

14. Two smooth oriented surfaces, $S_1$ and $S_2$ intersect in a piecewise smooth oriented closed curve, $C$. Let $\mathbf{F}$ be a $C^1$ vector field defined on $\mathbb{R}^3$. Explain why $\int_{S_1} \text{curl}(\mathbf{F}) \cdot \mathbf{n} \, d\mathbf{S} = \int_{S_2} \text{curl}(\mathbf{F}) \cdot \mathbf{n} \, d\mathbf{S}$. Here $\mathbf{n}$ is the normal to the surface which corresponds to the given orientation of the curve, $C$.

15. Show that $\text{curl}(\psi \nabla \phi) = \nabla \psi \times \nabla \phi$ and explain why $\int_{S} \nabla \psi \cdot \nabla \phi \cdot \mathbf{n} \, d\mathbf{S} = \int_{\partial S} (\psi \nabla \phi) \cdot \mathbf{n} \, d\mathbf{r}$.

16. Find a simple formula for $\text{div}(\nabla (u^\alpha))$ where $\alpha \in \mathbb{R}$.

17. Parametric equations for one arch of a cycloid are given by $x = a \left( t - \sin t \right)$ and $y = a \left( 1 - \cos t \right)$ where here $t \in [0, 2\pi]$. Sketch a rough graph of this arch of a cycloid and then find the area between this arch and the $x$ axis. Hint: This is very easy using Green’s theorem and the vector field, $\mathbf{F} = (-y, x)$.

18. Let $\mathbf{r}(t) = (\cos^3(t), \sin^3(t))$ where $t \in [0, 2\pi]$. Sketch this curve and find the area enclosed by it using Green’s theorem.

19. Consider the vector field, $\left( \frac{-y}{(x^2+y^2)}, \frac{x}{(x^2+y^2)}, 0 \right) = \mathbf{F}$. Show that $\nabla \times \mathbf{F} = \mathbf{0}$ but that for the closed curve, whose parameterization is $\mathbf{R}(t) = (\cos t, \sin t, 0)$ for $t \in [0, 2\pi]$, $\int_{C} \mathbf{F} \cdot d\mathbf{r} \neq 0$. Therefore, the vector field is not conservative. Does this contradict Theorem 24.2.11? Explain.

20. Let $\mathbf{x}$ be a point of $\mathbb{R}^3$ and let $\mathbf{n}$ be a unit vector. Let $D_r$ be the circular disk of radius $r$ containing $\mathbf{x}$ which is perpendicular to $\mathbf{n}$. Placing the tail of $\mathbf{n}$ at $\mathbf{x}$ and viewing $\partial D_r$, from the point of $\mathbf{n}$, orient $\partial D_r$ in the counter clockwise direction. Now suppose $\mathbf{F}$ is a vector field defined near $\mathbf{x}$. Show $\text{curl}(\mathbf{F}) \cdot \mathbf{n} = \lim_{r \to 0} \frac{1}{r^2} \int_{\partial D_r} \mathbf{F} \cdot d\mathbf{r}$. This last integral is sometimes called the circulation density of $\mathbf{F}$. Explain how this shows that $\text{curl}(\mathbf{F}) \cdot \mathbf{n}$ measures the tendency for the vector field to “curl” around the point, the vector $\mathbf{n}$ at the point $\mathbf{x}$.

21. The cylinder $x^2 + y^2 = 4$ is intersected with the plane $x + y + z = 2$. This yields a closed curve, $C$. Orient this curve in the counter clockwise direction when viewed from a point high on the $z$ axis. Let $\mathbf{F} = (x^2y, z + y, x^3)$. Find $\int_{C} \mathbf{F} \cdot d\mathbf{r}$.

22. The cylinder $x^2 + 4y^2 = 4$ is intersected with the plane $x + 3y + 2z = 1$. This yields a closed curve, $C$. Orient this curve in the counter clockwise direction when viewed from a point high on the $z$ axis. Let $\mathbf{F} = (y, z + y, x^2)$. Find $\int_{C} \mathbf{F} \cdot d\mathbf{r}$.

23. The cylinder $x^2 + y^2 = 4$ is intersected with the plane $x + 3y + 2z = 1$. This yields a closed curve, $C$. Orient this curve in the clockwise direction when viewed from a point high on the $z$ axis. Let $\mathbf{F} = (y, z + y, x)$. Find $\int_{C} \mathbf{F} \cdot d\mathbf{r}$.

24. Let $\mathbf{F} = (xz, z^2(y + \sin x), z^3y)$. Find the surface integral, $\int_{S} \text{curl}(\mathbf{F}) \cdot n \, dA$ where $S$ is the surface, $z = 4 - (x^2 + y^2)$, $z \geq 0$. 
25. Let \( \mathbf{F} = (xz, (y^3 + x), z^3y) \). Find the surface integral, \( \int \int_S \text{curl} (\mathbf{F}) \cdot \mathbf{n} \, dA \) where \( S \) is the surface, \( z = 16 - (x^2 + y^2) \), \( z \geq 0 \).

26. The cylinder \( z = y^2 \) intersects the surface \( z = 8 - x^2 - 4y^2 \) in a curve, \( C \) which is oriented in the counter clockwise direction when viewed high on the \( z \) axis. Find \( \int_C \mathbf{F} \cdot d\mathbf{R} \) if \( \mathbf{F} = \left( \frac{z}{2}, xy, xz \right) \). Hint: This is not too hard if you show you can use Stokes theorem on a domain in the \( xy \) plane.

27. Suppose solutions have been found to (24.17), (24.16), and (24.12). Then define \( \mathbf{E} \) and \( \mathbf{B} \) using (24.14) and (24.13). Verify Maxwell’s equations hold for \( \mathbf{E} \) and \( \mathbf{B} \).

28. Suppose now you have found solutions to (24.17) and (24.16), \( \psi_1 \) and \( A_1 \). Then go show again that if \( \phi \) satisfies (24.10) and \( \psi \equiv \psi_1 + \frac{1}{c} \frac{\partial \phi}{\partial t} \), while \( \mathbf{A} \equiv \mathbf{A}_1 + \nabla \phi \), then (24.12) holds for \( \mathbf{A} \) and \( \psi \).

29. Why consider Maxwell’s equations? Why not just consider (24.17), (24.16), and (24.12)?

30. Tell which open sets are simply connected.
   
   (a) The inside of a car radiator.
   (b) A donut.
   (c) The solid part of a cannon ball which contains a void on the interior.
   (d) The inside of a donut which has had a large bite taken out of it.
   (e) All of \( \mathbb{R}^3 \) except the \( z \) axis.
   (f) All of \( \mathbb{R}^3 \) except the \( xy \) plane.

31. Let \( P \) be a polygon with vertices \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), (x_1, y_1)\) encountered as you move over the boundary of the polygon in the counter clockwise direction. Using Problem 13, find a nice formula for the area of the polygon in terms of the vertices.
Curvilinear Coordinates

25.0.1 Outcomes

1. Define the basis and dual basis for general curvilinear coordinates.

2. Recall and use the transformation equations relating quantities in different curvilinear coordinate systems.

3. Define and use the Christoffel symbols.

4. Recall and use the formula for divergence and gradient in curvilinear coordinate systems.

You have already seen examples of curvilinear coordinates in the case of cylindrical and spherical coordinates. You use other coordinates other than \(x, y, z\) the rectangular coordinates to identify points. A general situation is the following: Let \(D \subseteq \mathbb{R}^n\) be an open set and let \(M : D \to \mathbb{R}^n\) satisfy

\[
\begin{align*}
M &\text{ is } C^2, \quad (25.1) \\
M &\text{ is one to one.} \quad (25.2)
\end{align*}
\]

Letting \(x \in D\), we can write

\[M(x) = M^k(x) i_k\]

where, \(i_k\) are the standard basis vectors for \(\mathbb{R}^n\), \(i_k\) being the vector in \(\mathbb{R}^n\) which has a one in the \(k^{th}\) coordinate and a 0 in every other spot. For a fixed \(x \in D\), we can consider the curves,

\[t \to M(x + ti_k)\]

for \(t \in I\), some open interval containing 0. Thus these curves are obtained by fixing all the variables except the \(k^{th}\) and then considering the curve which results. Then for the point \(x\),

\[e_k \equiv \frac{\partial M}{\partial x^k}(x)\]

Denote this vector as \(e_k(x)\) to emphasize its dependence on \(x\). The following picture illus-
It is desired that \( \{ e_k \}_{k=1}^n \) should be a basis. This holds if and only if
\[
\det \left( \frac{\partial M^i}{\partial x^k} \right) \neq 0. \tag{25.3}
\]
Let
\[
y^i = M^i(\mathbf{x}) \quad i = 1, \ldots, n \tag{25.4}
\]
so that the \( y^i \) are the usual coordinates with respect to the usual basis vectors \( \{ e_k \}_{k=1}^n \) of the point \( \mathbf{M}(\mathbf{x}) \). Letting \( \mathbf{x} \equiv (x^1, \ldots, x^n) \), it follows from the inverse function theorem of advanced calculus that \( \mathbf{M}(D) \) is open, and that (25.3), (25.1), and (25.2) imply the equations (25.4) define each \( x^i \) as a \( C^2 \) function of \( y \equiv (y^1, \ldots, y^n)^T \). Thus, abusing notation slightly, the equations (25.4) are equivalent to
\[
x^i = x^i(y), \quad i = 1, \ldots, n
\]
where \( x^i \) is a \( C^2 \) function. Thus
\[
\nabla x^k(y) = \frac{\partial x^k(y)}{\partial y^i} y^i.
\]
Then
\[
\nabla x^k(y) \cdot e_j = \frac{\partial x^k}{\partial y^s} y^s e_j = \frac{\partial x^k}{\partial y^s} \frac{\partial y^s}{\partial x^j} e_j = \delta^k_j
\]
by the chain rule. Therefore, the dual basis is given by
\[
e^k(x) = \nabla x^k(y). \tag{25.5}
\]
Notice that it might be hard or even impossible to solve algebraically for \( x^i \) in terms of the \( y^i \). Thus the straightforward approach to finding \( e^k \) by (25.5) might be impossible! Also, this approach leads to an expression in terms of the \( y \) coordinates rather than the desired \( x \) coordinates and so it is probably not a good idea to use it in the first place. It is expedient to use another method to obtain these vectors. The vectors, \( e^k(x) \) may always be found by raising the index using the inverse of the metric tensor as explained on Page 164 and the result is in terms of the curvilinear coordinates, \( x \). Consider the following example.

**Example 25.0.1** \( D \equiv (0, \infty) \times (0, \pi) \times (0, 2\pi) \) and
\[
\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} x^1 \sin(x^2) \cos(x^3) \\ x^1 \sin(x^2) \sin(x^3) \\ x^1 \cos(x^2) \end{pmatrix},
\]
usually written as

\[
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix} = \begin{pmatrix}
\rho \sin(\phi) \cos(\theta) \\
\rho \sin(\phi) \sin(\theta) \\
\rho \cos(\phi)
\end{pmatrix}
\]

where \((\rho, \phi, \theta)\) are the spherical coordinates. These coordinates are called \(x^1, x^2,\) and \(x^3\) to preserve the notation just discussed.) Thus

\[\mathbf{e}_1(x) = \sin(x^2) \cos(x^3) \mathbf{i}_1 + \sin(x^2) \sin(x^3) \mathbf{i}_2 + \cos(x^2) \mathbf{i}_3,\]

\[\mathbf{e}_2(x) = x^1 \cos(x^2) \cos(x^3) \mathbf{i}_1 + x^1 \cos(x^2) \sin(x^3) \mathbf{i}_2 - x^1 \sin(x^2) \mathbf{i}_3,\]

\[\mathbf{e}_3(x) = -x^1 \sin(x^2) \sin(x^3) \mathbf{i}_1 + x^1 \sin(x^2) \cos(x^3) \mathbf{i}_2 + 0 \mathbf{i}_3.\]

It follows the metric tensor is

\[
G = \begin{pmatrix}
1 & 0 & 0 \\
0 & (x^1)^2 & 0 \\
0 & 0 & (x^1)^2 \sin^2(x^2)
\end{pmatrix}
= (g_{ij}) = (\mathbf{e}_i \cdot \mathbf{e}_j). \tag{25.6}
\]

Therefore,

\[
G^{-1} = (g^{ij})
= (\mathbf{e}^i, \mathbf{e}^j) = \begin{pmatrix}
1 & 0 & 0 \\
0 & (x^1)^{-2} & 0 \\
0 & 0 & (x^1)^{-2} \sin^{-2}(x^2)
\end{pmatrix}.
\]

To obtain the dual basis, use Theorem 8.2.7 to write

\[
\mathbf{e}^1 = g^{1j} \mathbf{e}_j = \mathbf{e}_1
\]

\[
\mathbf{e}^2 = g^{2j} \mathbf{e}_j = (x^1)^{-2} \mathbf{e}_2
\]

\[
\mathbf{e}^3 = g^{3j} \mathbf{e}_j = (x^1)^{-2} \sin^{-2}(x^2) \mathbf{e}_3.
\]

It is natural to ask if there exists a transformation \(M\) such that

\[
\frac{\partial \mathbf{M}}{\partial x^3} = \mathbf{i} = \mathbf{i}_1, \frac{\partial \mathbf{M}}{\partial x^2} = \mathbf{j} = \mathbf{i}_2, \frac{\partial \mathbf{M}}{\partial x^1} = \mathbf{k} = \mathbf{i}_3. \tag{25.7}
\]

Let

\[
M(x^1, x^2, x^3) \equiv x^1 \mathbf{i} + x^2 \mathbf{j} + x^3 \mathbf{k}.
\]

Then (25.7) holds for this transformation.

**Example 25.0.2** Let

\[
\begin{pmatrix}
y^1 \\
y^2 \\
y^3
\end{pmatrix} = \begin{pmatrix}
3u + v \\
v - w \\
u - v^3
\end{pmatrix}
\]

where the \(y^i\) are the rectangular coordinates of the point. Find \(\mathbf{e}_i, i = 1, 2, 3,\) and find \((g_{ij})(x)\) and \((g^{ij}(x)).\)
First
\[ e_1 = \begin{pmatrix} 3 \\ 3 \\ 1 \end{pmatrix}, e_2 = \begin{pmatrix} 1 \\ 1 \\ -3v^2 \end{pmatrix}, e_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}. \]

Then the metric tensor is
\[
\begin{pmatrix} 3 & 0 & 1 \\ 1 & 1 & -3v^2 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 3 & 1 & 0 \\ 1 & 1 & -3v^2 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 10 & 3 - 3v^2 & 0 \\ 3 - 3v^2 & 2 + 9v^4 & 1 \\ 0 & 1 & 1 \end{pmatrix}
\]

Thus the inverse of the metric tensor is
\[
\begin{pmatrix} \frac{1+9v^4}{1+81v^2+18v^4} & \frac{v^2-1}{1+81v^2+18v^4} & \frac{v^2-1}{1+81v^2+18v^4} \\ \frac{3}{1+81v^2+18v^4} & \frac{3}{1+81v^2+18v^4} & \frac{3}{1+81v^2+18v^4} \\ \frac{9v^2-1}{1+81v^2+18v^4} & \frac{9v^2-1}{1+81v^2+18v^4} & \frac{9v^2-1}{1+81v^2+18v^4} \end{pmatrix}
\]

and so the dual basis consists of the columns of
\[
\begin{pmatrix} \frac{1+9v^4}{1+81v^2+18v^4} & \frac{3}{1+81v^2+18v^4} & \frac{3}{1+81v^2+18v^4} \\ \frac{3}{1+81v^2+18v^4} & \frac{3}{1+81v^2+18v^4} & \frac{3}{1+81v^2+18v^4} \\ \frac{9v^2-1}{1+81v^2+18v^4} & \frac{9v^2-1}{1+81v^2+18v^4} & \frac{9v^2-1}{1+81v^2+18v^4} \end{pmatrix}
\]

Clearly this is pretty horrible. That is because I picked a fairly arbitrary coordinate system.

### 25.1 Exercises

1. Let
\[
\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} x^1 + 2x^2 \\ x^2 + x^3 \\ x^1 - 2x^2 \end{pmatrix}
\]
where the \( y^i \) are the rectangular coordinates of the point. Find \( e^i, e_i, i = 1, 2, 3 \), and find \( (g_{ij}) (x) \) and \( (g^{ij}) (x) \).

2. Let
\[
\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} x^1 + x^2 \\ x^2 + x^3 \\ x^1 + 2x^2 \end{pmatrix}
\]
where the \( y^i \) are the rectangular coordinates of the point. Find \( e^i, e_i, i = 1, 2, 3 \), and find \( (g_{ij}) (x) \) and \( (g^{ij}) (x) \).

3. Let
\[
\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} x^1 + 2x^2 \\ x^3 \\ 3x^1 + x^3 \end{pmatrix}
\]
where the \( y^i \) are the rectangular coordinates of the point. Find \( e^i, e_i, i = 1, 2, 3 \), and find \( (g_{ij}) (x) \) and \( (g^{ij}) (x) \).
4. If the above are too easy and you want lots of computations to do, let
\[
\begin{pmatrix}
y_1 \\
y_2 \\
y_3
\end{pmatrix} = 
\begin{pmatrix}
x_1 + x^2 \\
x^2 + \sin(x^3) \\
x^1 + 2x^2
\end{pmatrix}
\]
where the \(y^i\) are the rectangular coordinates of the point. Find \(e^i, e_i, i = 1, 2, 3\), and find \((g_{ij})(x)\) and \((g^{ij})(x)\).

5. Let \(y = y(x,t)\) where \(t\) signifies time and \(x \in U \subseteq \mathbb{R}^m\) for \(U\) an open set, while \(y \in \mathbb{R}^n\) and suppose \(x\) is a function of \(t\). Physically, this corresponds to an object moving over a surface in \(\mathbb{R}^n\) which may be changing as a function of \(t\). The point \(y = y(x(t), t)\) is the point in \(\mathbb{R}^n\) corresponding to \(t\). For example, consider the pendulum

\[
\begin{array}{c}
\mu \\
\hline
l \\
\theta \\
m
\end{array}
\]

in which \(n = 2, l\) is fixed and \(y^1 = l \sin \theta, y^2 = l - l \cos \theta\). Thus, in this simple example, \(m = 1\). If \(l\) were changing in a known way with respect to \(t\), then this would be of the form \(y = y(x(t), t)\). The kinetic energy is defined as
\[
T \equiv \frac{1}{2} m \dot{y} \cdot \dot{y}
\]
where the dot on the top signifies differentiation with respect to \(t\). Show
\[
\frac{\partial T}{\partial y} = m \dot{y} \cdot \frac{\partial y}{\partial x^k}.
\]

Hint: First show
\[
\dot{y} = \frac{\partial y}{\partial x^j} \dot{x}^j + \frac{\partial y}{\partial t}
\]
and so
\[
\frac{\partial \dot{y}}{\partial x^j} = \frac{\partial y}{\partial x^j}.
\]

6. Show
\[
\frac{d}{dt} \left( \frac{\partial T}{\partial x^k} \right) = m \dot{y} \cdot \frac{\partial y}{\partial x^k} + m \dot{y} \cdot \frac{\partial^2 y}{\partial x^k \partial x^r} \dot{x}^r + m \dot{y} \cdot \frac{\partial^2 y}{\partial t \partial x^k}.
\]

7. Show
\[
\frac{\partial T}{\partial x^k} = m \dot{y} \cdot \left( \frac{\partial^2 y}{\partial x^r \partial x^k} \dot{x}^r + \frac{\partial^2 y}{\partial t \partial x^k} \right).
\]

Hint: Use * and **.

8. Now show from Newton’s second law (mass times acceleration equals force) that for \(F\) the force,
\[
\frac{d}{dt} \left( \frac{\partial T}{\partial x^k} \right) - \frac{\partial T}{\partial x^k} = m \dot{y} \cdot \frac{\partial y}{\partial x^k} = F \cdot \frac{\partial y}{\partial x^k}.
\]
9. In the example of the simple pendulum above,

\[ y = \left( \frac{l \sin \theta}{l - l \cos \theta} \right) = l \sin \theta \hat{i} + (l - l \cos \theta) \hat{j}. \]

Use ** to find a differential equation which describes the vibrations of the pendulum in terms of \( \theta \). First write the kinetic energy and then consider the force acting on the mass which is 

\[ -mg \hat{j}. \]

10. The above problem is fairly easy to do without the formalism developed. Now consider the case where \( \mathbf{x} = (\rho, \theta, \phi) \), spherical coordinates, and write differential equations for \( \rho, \theta, \) and \( \phi \) to describe the motion of an object in terms of these coordinates given a force, \( \mathbf{F} \).

11. Suppose the pendulum is not assumed to vibrate in a plane. Let it be suspended at the origin and consider spherical coordinates. Find differential equations for \( \theta \) and \( \phi \).

12. If there are many masses, \( m_\alpha, \alpha = 1, \cdots, R \), the kinetic energy is the sum of the kinetic energies of the individual masses. Thus,

\[ T \equiv \frac{1}{2} \sum_{\alpha=1}^{R} m_\alpha |\dot{y}_\alpha|^2. \]

Generalize the above problems to show that, assuming \( y_\alpha = y_\alpha(\mathbf{x}, t) \),

\[ \frac{d}{dt} \left( \frac{\partial T}{\partial \dot{x}^k} \right) = \frac{\partial T}{\partial x^k} = \sum_{\alpha=1}^{R} \mathbf{F}_\alpha \cdot \frac{\partial y_\alpha}{\partial x^k} \]

where \( \mathbf{F}_\alpha \) is the force acting on \( m_\alpha \).

13. Discuss the equivalence of these formulae with Newton’s second law, force equals mass times acceleration. What is gained from the above so called Lagrangian formalism?

14. The double pendulum has two masses instead of only one.

Write differential equations for \( \theta \) and \( \phi \) to describe the motion of the double pendulum.
25.2 Transformation Of Coordinates.

The scalars \( \{x^i\} \) are called curvilinear coordinates. Note they can be used to identify a point in \( \mathbb{R}^n \) and \( x = (x^1, \ldots, x^n) \) is a point in \( \mathbb{R}^n \). The basis vectors associated with this particular set of curvilinear coordinates at a point identified by \( \mathbf{x} \) are denoted by \( \mathbf{e}_i (\mathbf{x}) \) and the dual basis vectors at this point are denoted by \( \mathbf{e}^i (\mathbf{x}) \). What if other curvilinear coordinates are used? How do you write \( \mathbf{e}^k (\mathbf{x}) \) in terms of the vectors, \( \mathbf{e}^j (\mathbf{z}) \) where \( \mathbf{z} \) is some other type of curvilinear coordinates?

Consider the following picture in which \( U \) is an open set in \( \mathbb{R}^n \), \( D \), and \( \hat{D} \) are open sets in \( \mathbb{R}^n \), and \( \mathbf{M, N} \) are \( C^2 \) mappings which are one to one from \( D \) and \( \hat{D} \) respectively. Suppose that a point in \( U \) is identified by the curvilinear coordinates \( \mathbf{x} \) in \( D \) and \( \mathbf{z} \) in \( \hat{D} \).

![](image.png)

Thus \( \mathbf{M} (\mathbf{x}) = \mathbf{N} (\mathbf{z}) \). By the chain rule,

\[
e_i (\mathbf{z}) = \frac{\partial \mathbf{N}}{\partial x^i} = \frac{\partial \mathbf{M}}{\partial x^j} \frac{\partial x^j}{\partial z^i} = \frac{\partial x^i}{\partial z^i} \mathbf{e}_j (\mathbf{x}) \tag{25.8}
\]

Recall the covariant and contravariant coordinates defined in Chapter 8 which starts on Page 157. Thus,

\[
\mathbf{v} = v_i (\mathbf{x}) \mathbf{e}^i (\mathbf{x}) = v^j (\mathbf{x}) e_i (\mathbf{x}) = v_j (\mathbf{z}) \mathbf{e}^j (\mathbf{z}) = v^j (\mathbf{z}) \mathbf{e}_j (\mathbf{z}) .
\]

Then the following theorem tells how to transform various things defined above.

**Theorem 25.2.1** The following transformation rules hold for pairs of curvilinear coordinates.

\[
v_i (\mathbf{z}) = \frac{\partial x^j}{\partial z^i} v_j (\mathbf{x}) , \quad v^j (\mathbf{z}) = \frac{\partial x^i}{\partial z^j} v^j (\mathbf{x}) , \tag{25.9}
\]

\[
e_i (\mathbf{z}) = \frac{\partial x^j}{\partial z^i} e_j (\mathbf{x}) , \quad e^j (\mathbf{z}) = \frac{\partial z^i}{\partial x^j} e^j (\mathbf{x}) , \tag{25.10}
\]

\[
g_{ij} (\mathbf{z}) = \frac{\partial x^r}{\partial z^i} \frac{\partial x^s}{\partial z^j} g_{rs} (\mathbf{x}) , \quad g^{ij} (\mathbf{z}) = \frac{\partial z^i}{\partial x^r} \frac{\partial z^j}{\partial x^s} g^{rs} (\mathbf{x}) . \tag{25.11}
\]

**Proof:** The first part of (25.10) is shown in (25.8). Then, from (25.8) and Theorem 8.2.2 on Page 162

\[
e^j (\mathbf{z}) = e^j (\mathbf{z}) \cdot e_j (\mathbf{x}) e^i (\mathbf{x}) = e^j (\mathbf{z}) \cdot \frac{\partial z^k}{\partial x^j} \mathbf{e}_k (\mathbf{z}) e^i (\mathbf{x})
\]

and this proves the second part of (25.10). Now to show (25.9), use Theorem 8.2.2 on Page 162 again.

\[
v_i (\mathbf{z}) = \mathbf{v} \cdot e_i (\mathbf{z}) = \mathbf{v} \frac{\partial x^j}{\partial z^i} e_j (\mathbf{x}) = \frac{\partial x^j}{\partial z^i} v_j (\mathbf{x})
\]

and

\[
v^i (\mathbf{z}) = \mathbf{v} \cdot e^i (\mathbf{z}) = \mathbf{v} \frac{\partial z^i}{\partial x^j} e^j (\mathbf{x}) = \frac{\partial z^i}{\partial x^j} v^j (\mathbf{x}) .
\]
To verify (25.11),
\[ g_{ij}(z) = e_i(z) \cdot e_j(z) = e_r(x) \frac{\partial x^r}{\partial z^i} \cdot e_s(x) \frac{\partial x^s}{\partial z^j} = g_{rs}(x) \frac{\partial x^r}{\partial z^i} \frac{\partial x^s}{\partial z^j}. \]

This proves the theorem.

Denote by \( y \) the curvilinear coordinates with the property
\[ e_k(y) = i_k = e_k(y). \]

### 25.3 Differentiation And Christoffel Symbols

Let \( F: U \rightarrow \mathbb{R}^n \) be differentiable. \( F \) is a vector field and it is used to model force, velocity, acceleration, or any other vector quantity which may change from point to point in \( U \). Then
\[ \frac{\partial F(x)}{\partial x^j} \]
is a vector and so there exist scalars, \( F^i_j(x) \) and \( F_{i,j}(x) \) such that
\[ \frac{\partial F(x)}{\partial x^j} = F^i_j(x) e_i(x) = F_{i,j}(x) e^j(x). \]  \hspace{1cm} (25.12)

How do these scalars transform when the coordinates are changed?

**Theorem 25.3.1** If \( x \) and \( z \) are curvilinear coordinates,
\[ F^r_s(x) = F^i_j(x) \frac{\partial x^r}{\partial z^i} \frac{\partial z^j}{\partial x^s}, \quad F_{r,s}(x) \frac{\partial x^r}{\partial z^i} \frac{\partial z^j}{\partial x^s} = F_{i,j}(x). \]  \hspace{1cm} (25.13)

**Proof:**
\[ F^r_s(x) e_r(x) \equiv \frac{\partial F(x)}{\partial x^s} = \frac{\partial F(z)}{\partial z^j} \frac{\partial z^j}{\partial x^s} \equiv \]
\[ F^i_j(x) e_i(x) \frac{\partial z^j}{\partial x^s} = F^i_j(z) \frac{\partial x^r}{\partial z^i} \frac{\partial z^j}{\partial x^s} e_r(x) \]
which shows the first formula of (25.12). To show the other formula,
\[ F_{i,j}(z) e^i(z) \equiv \frac{\partial F(z)}{\partial z^i} = \frac{\partial F(x)}{\partial x^s} \frac{\partial x^s}{\partial z^j} \equiv \]
\[ F_{r,s}(x) e^r(x) \frac{\partial x^s}{\partial z^j} = F_{r,s}(x) \frac{\partial x^r}{\partial z^i} \frac{\partial z^j}{\partial x^s} e^j(z), \]
and this shows the second formula for transforming these scalars.

Now \( F(x) = F^i(x) e_i(x) \) and so by the product rule,
\[ \frac{\partial F}{\partial x^j} = \frac{\partial F^i}{\partial x^j} e_i(x) + F^i(x) \frac{\partial e_i(x)}{\partial x^j}. \]

Now \( \frac{\partial e_i(x)}{\partial x^j} \) is a vector and so there exist scalars, \( \{ k \}_{ij} \) such that
\[ \frac{\partial e_i(x)}{\partial x^j} = \{ k \}_{ij} e_k(x). \]

Therefore,
\[ \frac{\partial F}{\partial x^j} = \frac{\partial F^k}{\partial x^j} e_k(x) + F^i(x) \{ k \}_{ij} e_k(x). \]
25.3. DIFFERENTIATION AND CHRISTOFFEL SYMBOLS

which shows
\[ F^k_{ij}(x) = \frac{\partial F^k}{\partial x^j} + F^i(x) \begin{pmatrix} k \\ ij \end{pmatrix} + F^i_{ij}(x) \begin{pmatrix} k \\ ij \end{pmatrix} + F^i_{ik}(x) \begin{pmatrix} k \\ ij \end{pmatrix} \]

This is sometimes called the covariant derivative.

These scalars are called the Christoffel symbols of the second kind. The next theorem is devoted to properties of these Christoffel symbols. Before stating the theorem, recall that the mapping, \( M \), which defines the curvilinear coordinates is \( C^2 \). The reason for this is that it will be necessary to assert mixed partial derivatives are equal.

**Theorem 25.3.2** The Christoffel symbols of the second kind satisfy the following

\[ \frac{\partial e_i(x)}{\partial x^j} = \begin{pmatrix} k \\ ij \end{pmatrix} e_k(x), \quad (25.14) \]

\[ \frac{\partial e^i(x)}{\partial x^j} = -\begin{pmatrix} i \\ kj \end{pmatrix} e^k(x), \quad (25.15) \]

\[ \begin{pmatrix} k \\ ij \end{pmatrix} = \begin{pmatrix} k \\ ji \end{pmatrix}, \quad (25.16) \]

\[ \begin{pmatrix} m \\ ik \end{pmatrix} = \frac{g^{jm}}{2} \left[ \frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^i} - \frac{\partial g_{ik}}{\partial x^j} \right]. \quad (25.17) \]

**Proof:** Formula (25.14) is the definition of the Christoffel symbols. Consider (25.15) next. To do so, note

\[ e^i(x) \cdot e_k(x) = \delta^i_k. \]

Then from the product rule,

\[ \frac{\partial e^i(x)}{\partial x^j} \cdot e_k(x) + e^i(x) \cdot \frac{\partial e_k(x)}{\partial x^j} = 0. \]

Now from the definition,

\[ \frac{\partial e^i(x)}{\partial x^j} \cdot e_k(x) = -e^i(x) \begin{pmatrix} r \\ kj \end{pmatrix} e_r(x) = -\begin{pmatrix} i \\ k \end{pmatrix} e^k(x). \]

Therefore,

\[ \frac{\partial e^i(x)}{\partial x^j} = \left( \frac{\partial e^i(x)}{\partial x^k} \cdot e_k(x) \right) e^k(x) = -\begin{pmatrix} i \\ k \end{pmatrix} e^k(x). \]

This verifies (25.15).

Letting \( \frac{\partial M(x)}{\partial x^j} = e_j(x) \), it follows from equality of mixed partial derivatives,

\[ \begin{pmatrix} k \\ ij \end{pmatrix} e_k(x) = \frac{\partial e_i}{\partial x^j} = \frac{\partial^2 M}{\partial x^i \partial x^j} = \frac{\partial^2 M}{\partial x^k \partial x^j} = \frac{\partial e_j}{\partial x^k} = \begin{pmatrix} k \\ ji \end{pmatrix} e_k(x), \]

which shows (25.16).

It remains to show (25.17).

\[ \frac{\partial g_{ij}}{\partial x^k} = \frac{\partial e_i}{\partial x^k} \cdot e_j + e_i \cdot \frac{\partial e_j}{\partial x^k} = \begin{pmatrix} r \\ ik \end{pmatrix} e_r + e_j \cdot e_i \begin{pmatrix} r \\ jk \end{pmatrix}. \]

Therefore,

\[ \frac{\partial g_{ij}}{\partial x^k} = \begin{pmatrix} r \\ ik \end{pmatrix} g_{rj} + \begin{pmatrix} r \\ jk \end{pmatrix} g_{ri}. \quad (25.18) \]
Switching \( i \) and \( k \) while remembering (25.16) yields

\[
\frac{\partial g_{ij}}{\partial x^k} = \left\{ \frac{r}{ik} \right\} g_{rj} + \left\{ \frac{r}{ji} \right\} g_{rk}.
\]

(25.19)

Now switching \( j \) and \( k \) in (25.18),

\[
\frac{\partial g_{ik}}{\partial x^j} = \left\{ \frac{r}{ij} \right\} g_{rk} + \left\{ \frac{r}{jk} \right\} g_{ri}.
\]

(25.20)

Adding (25.18) to (25.19) and subtracting (25.20) yields

\[
\frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^i} - \frac{\partial g_{ik}}{\partial x^j} = 2 \left\{ \frac{r}{ik} \right\} g_{rj}.
\]

Now multiplying both sides by \( g^{jm} \) and using the fact that \((g^{ij})\) is the inverse matrix for \((g_{ij})\),

\[
2 \left\{ \frac{m}{ik} \right\} = g^{jm} \left( \frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^i} - \frac{\partial g_{ik}}{\partial x^j} \right)
\]

which proves (25.17).

This is a very interesting formula because it shows the Christoffel symbols are completely determined by the metric tensor and its derivatives.

### 25.4 Gradients And Divergence

It is very important to express the gradient and the divergence in general coordinate systems.

As before, \( y \) will denote the standard coordinates with respect to the usual basis vectors. Thus

\[
\mathbf{N}(y) \equiv y^k \mathbf{e}_k, \quad \mathbf{e}_k(y) = \mathbf{e}^k(y).
\]

Let \( \phi : U \to \mathbb{R} \) be a differentiable scalar function, sometimes called a "scalar field" in this subject. Write \( \phi(x) \) to denote the value of \( \phi \) at the point whose coordinates are \( x \). In general, it is convenient to follow this practice for any field, vector or scalar. Thus \( \mathbf{F}(x) \) is the value of a vector field at the point of \( U \) determined by the coordinates \( x \). (Remember, vectors are those things which are determined by direction and magnitude.) In the standard coordinates, the gradient is known. It is given by the following formula.

\[
\nabla \phi(y) = \frac{\partial \phi(y)}{\partial y^k} \mathbf{e}^k(y).
\]

Recall \( \mathbf{e}^k(y) = \mathbf{i}_k \), the vector whose Cartesian coordinates are all zero except for a 1 in the \( k \)th position. Therefore, using the chain rule, if the coordinates of the point of \( U \) are given as \( x \),

\[
\nabla \phi(x) = \nabla \phi(y)
\]

\[
= \frac{\partial \phi(x)}{\partial x^r} \frac{\partial y^k}{\partial x^r} \mathbf{e}^k(x) = \frac{\partial \phi(x)}{\partial x^r} \delta^r_s \mathbf{e}^s(x) = \frac{\partial \phi(x)}{\partial x^r} \mathbf{e}^r(x).
\]

This shows the covariant components of \( \nabla \phi(x) \) are

\[
(\nabla \phi(x))_r = \frac{\partial \phi(x)}{\partial x^r}.
\]

(25.21)
To find the contravariant components, raise the index in the usual way. Thus

\[
(\nabla \phi (x))^r = g^{rk} (x) (\nabla \phi (x))_k = g^{rk} (x) \frac{\partial \phi (x)}{\partial x^k}.
\] (25.22)

What about the divergence of a vector field? The divergence of a vector field, \( \mathbf{F} \) defined on \( U \) is a scalar field, \( \text{div} (\mathbf{F}) \) which from calculus is

\[
\frac{\partial F^k}{\partial y^k} (y) = F^k (y)
\]

in terms of the usual coordinates \( y \). The reason the above equation holds in this case is that \( e_k (y) \) is a constant and so the Christoffel symbols are zero. What is the expression for the divergence in an arbitrary coordinate system? From Theorem 25.3.1,

\[
F^i_j (y) = F^r_s (x) \frac{\partial x^s}{\partial y^i} \frac{\partial y^j}{\partial x^r}
\]

Letting \( j = i \) yields

\[
\text{div} (\mathbf{F}) = \left( \frac{\partial F^r}{\partial x^s} + F^k (x) \left\{ \begin{array}{c} r \\ ks \end{array} \right\} (x) \right) \frac{\partial x^s}{\partial y^i} \frac{\partial y^i}{\partial x^r}
\]

(25.23)

The symbol, \( \left\{ \begin{array}{c} r \\ kr \end{array} \right\} \) is now simplified\(^1\) using the description of it in Theorem 25.3.2. Thus, from this theorem,

\[
\left\{ \begin{array}{c} r \\ k \end{array} \right\} = g^{jr} \left[ \frac{\partial g_{jr}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^r} - \frac{\partial g_{rk}}{\partial x^j} \right]
\]

Now consider \( \frac{g^{jr}}{2} \) times the last two terms in \( [\cdot] \). Relabeling the indices \( r \) and \( j \) in the second term implies

\[
\frac{g^{jr}}{2} \frac{\partial g_{kj}}{\partial x^r} - \frac{g^{jr}}{2} \frac{\partial g_{jk}}{\partial x^r} = \frac{g^{jr}}{2} \frac{\partial g_{kj}}{\partial x^r} - \frac{g^{jr}}{2} \frac{\partial g_{jk}}{\partial x^r} = 0.
\]

Therefore,

\[
\left\{ \begin{array}{c} r \\ rk \end{array} \right\} = \frac{g^{jr}}{2} \frac{\partial g_{rj}}{\partial x^k}.
\] (25.24)

Now recall \( g \equiv \det (g_{ij}) = \det (G) > 0 \) from Theorem 8.2.7 on Page 164. Also from the formula for the inverse of a matrix in which the inverse equals one divided by the determinant times the transpose of the cofactor matrix, and this theorem,

\[
g^{jr} = A^{jr} (\det G)^{-1} = A^{jr} (\det G)^{-1}
\]

where \( A^{jr} \) is the \( r^j \)th cofactor of the matrix \( (g_{ij}) \). Also recall that

\[
g = \sum_{r=1}^{n} g_{rj} A^{jr} \text{ no sum on } j.
\]

\(^1\)This is called a contraction because there is a repeated index in the symbol.
Therefore, $g$ is a function of the variables \{${g_{rj}}$\} and
\[
\frac{\partial g}{\partial g_{rj}} = A_{rj}.
\]
From (25.24),
\[
\left\{ \begin{array}{l}
 r \\
 r k
\end{array} \right\} = g^{r j} \frac{\partial g_{rj}}{\partial x^k} = \frac{1}{2g} \frac{\partial g_{rj}}{\partial x^k} A^r = \frac{1}{2g} \frac{\partial g}{\partial g_{rj}} = \frac{1}{2g} \frac{\partial g}{\partial x^k}
\]
and so from (25.23),
\[
\text{div} (\mathbf{F}) = \frac{\partial F_k}{\partial x^k} + F_k \frac{1}{2g} \frac{\partial g}{\partial x^k} = \frac{1}{\sqrt{g}} \frac{\partial}{\partial x^i} \left( F^i \sqrt{g} \right).
\]
This is the formula for the divergence of a vector field in general curvilinear coordinates.

The Laplacian of a scalar field is nothing more than the divergence of the gradient. In symbols,
\[
\Delta \phi \equiv \nabla \cdot \nabla \phi
\]
From (25.25) and (25.22) it follows
\[
\Delta \phi (x) = \frac{1}{\sqrt{g}} \frac{\partial}{\partial x^i} \left( g^{ik} \frac{\partial \phi}{\partial x^k} \sqrt{g} \right).
\]

To summarize the conclusions of this section, here is a major theorem.

**Theorem 25.4.1** The following formulas hold for the gradient, divergence and Laplacian in general curvilinear coordinates.
\[
(\nabla \phi) (x) = \frac{\partial \phi}{\partial x^r},
\]
\[
(\nabla \phi) (x) = g^{r k} \frac{\partial \phi}{\partial x^k},
\]
\[
\text{div} (\mathbf{F}) = \frac{1}{\sqrt{g}} \frac{\partial}{\partial x^i} \left( F^i \sqrt{g} \right),
\]
\[
\Delta \phi (x) = \frac{1}{\sqrt{g}} \frac{\partial}{\partial x^i} \left( g^{ik} \frac{\partial \phi}{\partial x^k} \sqrt{g} \right).
\]

### 25.5 Exercises

1. Let $y^1 = x^1 + 2x^2, y^2 = x^2 + 3x^3, y^3 = x^1 + x^3$. Let
\[
\mathbf{F} (x) = x^1 \mathbf{e}_1 (x) + x^2 \mathbf{e}_2 (x) + (x^3)^2 \mathbf{e} (x).
\]
Find $\text{div} (\mathbf{F}) (x)$.

2. For the coordinates of the preceding problem, and $\phi$ a scalar field, find
\[
(\nabla \phi (x))^3
\]
in terms of the partial derivatives of $\phi$ taken with respect to the variables $x^i$.
3. Let \( y^1 = 7x^1 + 2x^2, y^2 = x^2 + 3x^3, y^3 = x^1 + x^3 \). Let \( \phi \) be a scalar field. Find \( \nabla^2 \phi(x) \).

4. Derive \( \nabla^2 u \) in cylindrical coordinates, \( r, \theta, z \), where \( u \) is a scalar field on \( \mathbb{R}^3 \).

\[
\begin{align*}
x &= r \cos \theta, & y &= r \sin \theta, & z &= z. 
\end{align*}
\]

5. Find all solutions to \( \nabla^2 u = 0 \) which depend only on \( r \) where \( r \equiv \sqrt{x^2 + y^2} \).

6. Let \( u \) be a scalar field on \( \mathbb{R}^3 \). Find all solutions to \( \nabla^2 u = 0 \) which depend only on

\[
\rho \equiv \sqrt{x^2 + y^2 + z^2}.
\]

7. Find all harmonic functions defined on \( \mathbb{R}^n \) which depend only in \( \rho \), the distance to the origin. Hint: Use the formula for the Laplacian in an appropriate coordinate system.

8. The temperature, \( u \), in a solid satisfies \( \nabla^2 u = 0 \) after a long time. Suppose in a long pipe of inner radius 9 and outer radius 10 the exterior surface is held at 100\(^\circ\) while the inner surface is held at 200\(^\circ\) find the temperature in the solid part of the pipe.

9. Show

\[
\left\{ \frac{l}{ij} \right\} = \frac{\partial e_i}{\partial x^j} \cdot e^l.
\]

Find the Christoffel symbols of the second kind for spherical coordinates in which \( x^1 = \phi, x^2 = \theta, \) and \( x^3 = \rho \). Do the same for cylindrical coordinates letting \( x^1 = r, x^2 = \theta, x^3 = z \).

10. Show velocity can be expressed as \( v = v_i(x) e^i(x) \), where

\[
v_i(x) = \frac{\partial r_i}{\partial x^j} \frac{dx^j}{dt} - r_p(x) \left\{ \frac{p}{ik} \right\} \frac{dx^k}{dt}
\]

and \( r_i(x) \) are the covariant components of the displacement vector,

\[
r = r_i(x) e^i(x).
\]

11. Using problem 9 and 10, show the covariant components of velocity in spherical coordinates are

\[
v_1 = \rho^2 \frac{d\phi}{dt}, \quad v_2 = \rho^2 \sin^2(\phi) \frac{d\theta}{dt}, \quad v_3 = \frac{d\rho}{dt}
\]

Hint: First observe that if \( r \) is the position vector from the origin, then \( r = \rho e_3 \) so \( r_1 = 0 = r_2 \), and \( r_3 = \rho \). Now use 10.

\[25.6 \quad \text{Curl And Cross Products}\]

I have not had occasion to use this material very much but for the sake of completeness, here it is. It involves the curl and cross product in terms of general curvilinear coordinates. It will always be assumed that for \( x \) a set of curvilinear coordinates,

\[
\det \left( \frac{\partial y^i}{\partial x^j} \right) > 0 \quad (25.31)
\]

Where the \( y_i \) are the usual coordinates in which \( e_k(y) = i_k \). This sort of fussy thing is necessary because of the antisymmetry of the cross product.
**Theorem 25.6.1** Let (25.31) hold. Then

\[
\det \left( \frac{\partial y^i}{\partial x^j} \right) = \sqrt{g(x)} \quad (25.32)
\]

and

\[
\det \left( \frac{\partial x^i}{\partial y^j} \right) = \frac{1}{\sqrt{g(x)}} \quad (25.33)
\]

**Proof:**

\[e_i(x) = \frac{\partial y^k}{\partial x^i} i_k\]

and so

\[g_{ij}(x) = \frac{\partial y^k}{\partial x^i} i_k \cdot \frac{\partial y^l}{\partial x^j} i_l = \frac{\partial y^k}{\partial x^i} \frac{\partial y^k}{\partial x^j}.\]

Therefore, \[g = \det (g_{ij}(x)) = \left( \det \left( \frac{\partial y^k}{\partial x^i} \right) \right)^2.\] By (25.31), \[\sqrt{g} = \det \left( \frac{\partial y^k}{\partial x^i} \right)\] as claimed. Now

\[\frac{\partial y^k}{\partial x^i} \frac{\partial x^i}{\partial y^r} = \delta^k_r\]

and so

\[\det \left( \frac{\partial x^i}{\partial y^r} \right) = \frac{1}{\sqrt{g(x)}}\]

This proves the theorem.

To get the curl and cross product in curvilinear coordinates, let \(\epsilon^{ijk}\) be the usual permutation symbol. Thus,

\[\epsilon^{123} = 1\]

and when any two indices in \(\epsilon^{ijk}\) are switched, the sign changes. Thus

\[\epsilon^{132} = -1, \epsilon^{312} = 1, \text{ etc.}\]

Now define

\[\epsilon^{ijk}(x) \equiv \epsilon^{ijk} \frac{1}{\sqrt{g(x)}}\]

Then for \(x\) and \(z\) satisfying (25.31),

\[\epsilon^{ijk}(x) \frac{\partial z^r}{\partial x^i} \frac{\partial z^s}{\partial x^j} \frac{\partial z^t}{\partial x^k} = \epsilon^{ijk} \det \left( \frac{\partial y^p}{\partial y^q} \right) \frac{\partial z^r}{\partial x^i} \frac{\partial z^s}{\partial x^j} \frac{\partial z^t}{\partial x^k} = \epsilon^{rst} \det (MN)\]

where \(N\) is the matrix whose \(pq\text{th}\) entry is \(\frac{\partial y^p}{\partial x^q}\) and \(M\) is the matrix whose \(ik\text{th}\) entry is \(\frac{\partial z^i}{\partial x^k}\). Therefore, from the definition of matrix multiplication and the chain rule, this equals

\[\epsilon^{rst} \det \left( \frac{\partial x^p}{\partial y^q} \right) = \epsilon^{rst} \det (z)\]

from the above discussion.

Now \(\epsilon^{ijk}(y) = \epsilon^{ijk}\) and for a vector field, \(F\),

\[\text{curl } (F) \equiv \epsilon^{ijk}(y) F_{k,j}(y) e_i(y).\]
Therefore, since we know how everything transforms assuming (25.31), it is routine to write this in terms of \( x \).

\[
\text{curl} (\mathbf{F}) = \varepsilon^{rst} (x) \frac{\partial y^i}{\partial x^r} \frac{\partial y^j}{\partial x^s} \frac{\partial y^k}{\partial x^t} F_{p,q} (x) \frac{\partial x^p}{\partial y^q} \frac{\partial x^q}{\partial y^p} \mathbf{e}_m (x) \frac{\partial x^m}{\partial y^i} \\
= \varepsilon^{rst} (x) \delta_3^m \delta_3^p \delta_3^q F_{p,q} (x) \mathbf{e}_m (x) = \varepsilon^{mpq} (x) F_{p,q} (x) \mathbf{e}_m (x). \tag{25.34}
\]

More simplification is possible. Recalling the definition of \( F_{p,q} (x) \),

\[
\frac{\partial \mathbf{F}}{\partial x^q} \equiv F_{p,q} (x) \mathbf{e}^p (x) = \frac{\partial}{\partial x^q} [F_p (x) \mathbf{e}^p (x)]
\]

\[
= \frac{\partial F_p (x)}{\partial x^q} \mathbf{e}^p (x) + F_p (x) \frac{\partial \mathbf{e}^p}{\partial x^q} = \frac{\partial F_p (x)}{\partial x^q} \mathbf{e}^p (x) - F_r (x) \left\{ \frac{r}{pq} \right\} \mathbf{e}^p (x)
\]

by Theorem 25.3.2. Therefore,

\[
F_{p,q} (x) = \frac{\partial F_p (x)}{\partial x^q} - F_r (x) \left\{ \frac{r}{pq} \right\}
\]

and so

\[
\text{curl} (\mathbf{F}) = \varepsilon^{mpq} (x) \frac{\partial F_p (x)}{\partial x^q} \mathbf{e}_m (x) - \varepsilon^{mpq} (x) F_r (x) \left\{ \frac{r}{pq} \right\} \mathbf{e}_m (x).
\]

However, because \( \left\{ \frac{r}{pq} \right\} = \left\{ \frac{r}{qp} \right\} \), the second term in this expression equals 0. To see this,

\[
\varepsilon^{mpq} (x) \left\{ \frac{r}{pq} \right\} = \varepsilon^{mpq} (x) \left\{ \frac{r}{qp} \right\} = -\varepsilon^{mpq} (x) \left\{ \frac{r}{pq} \right\}.
\]

Therefore, by (25.34),

\[
\text{curl} (\mathbf{F}) = \varepsilon^{mpq} (x) \frac{\partial F_p (x)}{\partial x^q} \mathbf{e}_m (x). \tag{25.35}
\]

What about the cross product of two vector fields? Let \( \mathbf{F} \) and \( \mathbf{G} \) be two vector fields. Then in terms of standard coordinates, \( y \),

\[
\mathbf{F} \times \mathbf{G} = \varepsilon^{ijk} (y) F_j (y) G_k (y) \mathbf{e}_i (y)
\]

\[
= \varepsilon^{rst} (x) \frac{\partial y^i}{\partial x^r} \frac{\partial y^j}{\partial x^s} \frac{\partial y^k}{\partial x^t} F_p (x) \frac{\partial x^p}{\partial y^q} G_q (x) \frac{\partial x^q}{\partial y^p} \mathbf{e}_i (x) \frac{\partial x^l}{\partial y^q}
\]

\[
= \varepsilon^{rst} (x) \delta_3^i \delta_3^j \delta_3^k F_p (x) G_q (x) \mathbf{e}_l (x) = \varepsilon^{ipq} (x) F_p (x) G_q (x) \mathbf{e}_l (x). \tag{25.36}
\]

The above is summarized in the following theorem.

**Theorem 25.6.2** Suppose \( x \) is a system of curvilinear coordinates in \( \mathbb{R}^3 \) such that

\[
\det \left( \frac{\partial y^i}{\partial x^j} \right) > 0.
\]

Let

\[
\varepsilon^{ijk} (x) \equiv \varepsilon^{ijk} \frac{1}{\sqrt{g (x)}}.
\]

Then the following formulas for curl and cross product hold in this system of coordinates.

\[
\text{curl} (\mathbf{F}) = \varepsilon^{mpq} (x) \frac{\partial F_p (x)}{\partial x^q} \mathbf{e}_m (x),
\]

and

\[
\mathbf{F} \times \mathbf{G} = \varepsilon^{ipq} (x) F_p (x) G_q (x) \mathbf{e}_l (x).
\]
The Theory Of The Riemann Integral

The definition of the Riemann integral of a function of \( n \) variables uses the following definition.

**Definition 26.0.3** For \( i = 1, \ldots, n \), let \( \{a_i^k\}_{k=\infty}^{-\infty} \) be points on \( \mathbb{R} \) which satisfy

\[
\lim_{k \to \infty} a_i^k = \infty, \quad \lim_{k \to -\infty} a_i^k = -\infty, \quad a_i^k < a_i^{k+1}.
\]  

(26.1)

For such sequences, define a grid on \( \mathbb{R}^n \) denoted by \( G \) or \( F \) as the collection of boxes of the form

\[
Q = \prod_{i=1}^{n} [a_i^j, a_i^{j+1}].
\]

(26.2)

If \( G \) is a grid, \( F \) is called a refinement of \( G \) if every box of \( G \) is the union of boxes of \( F \).

**Lemma 26.0.4** If \( G \) and \( F \) are two grids, they have a common refinement, denoted here by \( G \lor F \).

**Proof:** Let \( \{a_i^k\}_{k=\infty}^{-\infty} \) be the sequences used to construct \( G \) and let \( \{b_i^k\}_{k=\infty}^{-\infty} \) be the sequence used to construct \( F \). Now let \( \{\gamma_i^k\}_{k=\infty}^{-\infty} \) denote the union of \( \{a_i^k\}_{k=\infty}^{-\infty} \) and \( \{b_i^k\}_{k=\infty}^{-\infty} \). It is necessary to show that for each \( i \) these points can be arranged in order. To do so, let \( \gamma_0^i \equiv a_0^i \). Now if

\[
\gamma_{i-j}^1, \ldots, \gamma_0^i, \ldots, \gamma_j^i
\]

have been chosen such that they are in order and all distinct, let \( \gamma_{j+1}^i \) be the first element of

\[
\{a_i^k\}_{k=\infty}^{-\infty} \cup \{b_i^k\}_{k=\infty}^{-\infty}
\]

(26.3)
which is larger than $\gamma_{i,j}^j$ and let $\gamma_{i,(j+1)}^i$ be the last element of (26.3) which is strictly smaller than $\gamma_{i,j}^i$. The assumption (20.8) insures such a first and last element exists. Now let the grid $\mathcal{G} \cup \mathcal{F}$ consist of boxes of the form

$$Q = \prod_{i=1}^{n} [\gamma_{i,j}^i, \gamma_{i,j+1}^i].$$

The Riemann integral is only defined for functions, $f$ which are bounded and are equal to zero out of some bounded set, $D$. In what follows $f$ will always be such a function.

**Definition 26.0.5** Let $f$ be a bounded function which equals zero off a bounded set, $D$, and let $\mathcal{G}$ be a grid. For $Q \in \mathcal{G}$, define

$$M_Q (f) \equiv \sup \{ f(x) : x \in Q \}, \quad m_Q (f) \equiv \inf \{ f(x) : x \in Q \}.$$  \hfill (26.4)

Also define for $Q$ a box, the volume of $Q$, denoted by $v(Q)$ by

$$v(Q) \equiv \prod_{i=1}^{n} (b_i - a_i), \quad Q \equiv \prod_{i=1}^{n} [a_i, b_i].$$

Now define upper sums, $U_{\mathcal{G}} (f)$ and lower sums, $L_{\mathcal{G}} (f)$ with respect to the indicated grid, by the formulas

$$U_{\mathcal{G}} (f) \equiv \sum_{Q \in \mathcal{G}} M_Q (f) v(Q), \quad L_{\mathcal{G}} (f) \equiv \sum_{Q \in \mathcal{G}} m_Q (f) v(Q).$$

A function of $n$ variables is Riemann integrable when there is a unique number between all the upper and lower sums. This number is the value of the integral.

Note that in this definition, $M_Q (f) = m_Q (f) = 0$ for all but finitely many $Q \in \mathcal{G}$ so there are no convergence questions to be considered here.

**Lemma 26.0.6** If $\mathcal{F}$ is a refinement of $\mathcal{G}$ then

$$U_{\mathcal{G}} (f) \geq U_{\mathcal{F}} (f), \quad L_{\mathcal{G}} (f) \leq L_{\mathcal{F}} (f).$$

Also if $\mathcal{F}$ and $\mathcal{G}$ are two grids,

$$L_{\mathcal{G}} (f) \leq U_{\mathcal{F}} (f).$$

**Proof:** For $P \in \mathcal{G}$ let $\hat{P}$ denote the set,

$$\{ Q \in \mathcal{F} : Q \subseteq P \}.$$

Then $P = \cup \hat{P}$ and

$$L_{\mathcal{F}} (f) \equiv \sum_{Q \in \mathcal{F}} m_Q (f) v(Q) = \sum_{P \in \hat{P}} \sum_{Q \in P} m_Q (f) v(Q)$$

$$\geq \sum_{P \in \hat{P}} m_P (f) \sum_{Q \in \hat{P}} v(Q) = \sum_{P \in \hat{P}} m_P (f) v(P) \equiv L_{\mathcal{G}} (f).$$

Similarly, the other inequality for the upper sums is valid.

To verify the last assertion of the lemma, use Lemma 26.0.4 to write

$$L_{\mathcal{G}} (f) \leq L_{\mathcal{G} \cup \mathcal{F}} (f) \leq U_{\mathcal{G} \cup \mathcal{F}} (f) \leq U_{\mathcal{F}} (f).$$

This proves the lemma.

This lemma makes it possible to define the Riemann integral.
Definition 26.0.7 Define an upper and a lower integral as follows.

\[ T(f) \equiv \inf \{ U_G(f) : G \text{ is a grid} \} , \]
\[ I(f) \equiv \sup \{ L_G(f) : G \text{ is a grid} \} . \]

Lemma 26.0.8 \( T(f) \geq I(f) \).

Proof: From Lemma 26.0.6 it follows for any two grids \( G \) and \( F \),

\[ L_G(f) \leq U_F(f) . \]

Therefore, taking the supremum for all grids on the left in this inequality,

\[ I(f) \leq U_F(f) \]

for all grids \( F \). Taking the infimum in this inequality, yields the conclusion of the lemma.

Definition 26.0.9 A bounded function, \( f \) which equals zero off a bounded set, \( D \), is said to be Riemann integrable, written as \( f \in R(\mathbb{R}^n) \) exactly when \( I(f) = T(f) \). In this case define

\[ \int f \, dV \equiv \int f \, dx = I(f) = T(f) . \]

As in the case of integration of functions of one variable, one obtains the Riemann criterion which is stated as the following theorem.

Theorem 26.0.10 (Riemann criterion) \( f \in R(\mathbb{R}^n) \) if and only if for all \( \varepsilon > 0 \) there exists a grid \( G \) such that

\[ U_G(f) - L_G(f) < \varepsilon . \]

Proof: If \( f \in R(\mathbb{R}^n) \), then \( T(f) = I(f) \) and so there exist grids \( G \) and \( F \) such that

\[ U_G(f) - L_F(f) \leq T(f) + \frac{\varepsilon}{2} - (I(f) - \frac{\varepsilon}{2}) = \varepsilon . \]

Then letting \( H = G \lor F \), Lemma 26.0.6 implies

\[ U_H(f) - L_H(f) \leq U_G(f) - L_F(f) < \varepsilon . \]

Conversely, if for all \( \varepsilon > 0 \) there exists \( G \) such that

\[ U_H(f) - L_H(f) < \varepsilon , \]

then

\[ T(f) - I(f) \leq U_G(f) - L_G(f) < \varepsilon . \]

Since \( \varepsilon > 0 \) is arbitrary, this proves the theorem.

26.1 Basic Properties

It is important to know that certain combinations of Riemann integrable functions are Riemann integrable. The following theorem will include all the important cases.
Theorem 26.1.1 Let \( f, g \in \mathcal{R}(\mathbb{R}^n) \) and let \( \phi : K \to \mathbb{R} \) be continuous where \( K \) is a compact set in \( \mathbb{R}^2 \) containing \( f(\mathbb{R}^n) \times g(\mathbb{R}^n) \). Also suppose that \( \phi(0,0) = 0 \). Then defining

\[
h(x) \equiv \phi(f(x), g(x)),
\]

it follows that \( h \) is also in \( \mathcal{R}(\mathbb{R}^n) \).

**Proof:** Let \( \varepsilon > 0 \) and let \( \delta_1 > 0 \) be such that if \( (y_i, z_i), i = 1, 2 \) are points in \( K \), such that \( |z_1 - z_2| \leq \delta_1 \) and \( |y_1 - y_2| \leq \delta_1 \), then

\[
|\phi(y_1, z_1) - \phi(y_2, z_2)| < \varepsilon.
\]

Let \( 0 < \delta < \min(\delta_1, \varepsilon, 1) \). Let \( \mathcal{G} \) be a grid with the property that for \( Q \in \mathcal{G} \), the diameter of \( Q \) is less than \( \delta \) and also for \( k = f, g \),

\[
U_{\mathcal{G}}(k) - L_{\mathcal{G}}(k) < \delta^2. \tag{26.5}
\]

Then defining for \( k = f, g \),

\[
P_k \equiv \{ Q \in \mathcal{G} : M_Q(k) - m_Q(k) > \delta \},
\]

it follows

\[
\delta^2 > \sum_{Q \in \mathcal{G}} (M_Q(k) - m_Q(k)) v(Q) \geq \sum_{P_k} (M_Q(k) - m_Q(k)) v(Q) \geq \delta \sum_{P_k} v(Q)
\]

and so for \( k = f, g \),

\[
\varepsilon > \delta > \sum_{P_k} v(Q). \tag{26.6}
\]

Suppose for \( k = f, g \),

\[
M_Q(k) - m_Q(k) \leq \delta.
\]

Then if \( x_1, x_2 \in Q \),

\[
|f(x_1) - f(x_2)| < \delta, \text{ and } |g(x_1) - g(x_2)| < \delta.
\]

Therefore,

\[
|h(x_1) - h(x_2)| \equiv |\phi(f(x_1), g(x_1)) - \phi(f(x_2), g(x_2))| < \varepsilon
\]

and it follows that

\[
|M_Q(h) - m_Q(h)| \leq \varepsilon.
\]

Now let

\[
\mathcal{S} \equiv \{ Q \in \mathcal{G} : 0 < M_Q(k) - m_Q(k) \leq \delta, \ k = f, g \}.
\]

Thus the union of the boxes in \( \mathcal{S} \) is contained in some large box, \( R \), which depends only on \( f \) and \( g \) and also, from the assumption that \( \phi(0,0) = 0 \), \( M_Q(h) - m_Q(h) = 0 \) unless \( Q \subseteq R \). Then

\[
U_{\mathcal{G}}(h) - L_{\mathcal{G}}(h) \leq \sum_{Q \in \mathcal{F}_f} (M_Q(h) - m_Q(h)) v(Q) + \sum_{Q \in \mathcal{F}_g} (M_Q(h) - m_Q(h)) v(Q) + \sum_{Q \in \mathcal{S}} \delta v(Q).
\]
Now since $K$ is compact, it follows $\phi(K)$ is bounded and so there exists a constant, $C$, depending only on $h$ and $\phi$ such that $M_Q(h) - m_Q(h) < C$. Therefore, the above inequality implies
\[
U_G(h) - L_G(h) \leq C \sum_{Q \in \mathcal{P}_\delta} v(Q) + C \sum_{Q \in \mathcal{P}_g} v(Q) + \sum_{Q \in \mathcal{S}} \delta v(Q),
\]
which by (26.6) implies
\[
U_G(h) - L_G(h) \leq 2C \varepsilon + \delta v(R) \leq 2C \varepsilon + \varepsilon v(R).
\]
Since $\varepsilon$ is arbitrary, the Riemann criterion is satisfied and so $h \in \mathcal{R}(\mathbb{R}^n)$.

**Corollary 26.1.2** Let $f, g \in \mathcal{R}(\mathbb{R}^n)$ and let $a, b \in \mathbb{R}$. Then $af + bg$, $fg$, and $|f|$ are all in $\mathcal{R}(\mathbb{R}^n)$. Also,
\[
\int_{\mathbb{R}^n} (af + bg) \, dx = a \int_{\mathbb{R}^n} f \, dx + b \int_{\mathbb{R}^n} g \, dx,
\]
and
\[
\int |f| \, dx \geq \left| \int f \, dx \right|.
\]

**Proof:** Each of the combinations of functions described above is Riemann integrable by Theorem 26.1.1. For example, to see $af + bg \in \mathcal{R}(\mathbb{R}^n)$ consider $\phi(y, z) \equiv ay + bz$. This is clearly a continuous function of $(y, z)$ such that $\phi(0, 0) = 0$. To obtain $|f| \in \mathcal{R}(\mathbb{R}^n)$, let $\phi(y, z) \equiv |y|$. It remains to verify the formulas. To do so, let $G$ be a grid with the property that for $k = f, g, |f|$ and $af + bg$,
\[
U_G(k) - L_G(k) < \varepsilon.
\]

Consider (26.7). For each $Q \in G$ pick a point in $Q, x_Q$. Then
\[
\sum_{Q \in G} k(x_Q) v(Q) \in [L_G(k), U_G(k)]
\]
and so
\[
\left| \int k \, dx - \sum_{Q \in G} k(x_Q) v(Q) \right| < \varepsilon.
\]
Consequently, since
\[
\sum_{Q \in G} (af + bg)(x_Q) v(Q) = a \sum_{Q \in G} f(x_Q) v(Q) + b \sum_{Q \in G} g(x_Q) v(Q),
\]
it follows
\[
\left| \int (af + bg) \, dx - a \int f \, dx - b \int g \, dx \right| \leq \left| \int (af + bg) \, dx - \sum_{Q \in G} (af + bg)(x_Q) v(Q) \right| + \left| a \sum_{Q \in G} f(x_Q) v(Q) - a \int f \, dx \right| + \left| b \sum_{Q \in G} g(x_Q) v(Q) - b \int g \, dx \right|,
\]
$\leq \varepsilon + |a| \varepsilon + |b| \varepsilon.$

Since $\varepsilon$ is arbitrary, this establishes Formula (26.7) and shows the integral is linear.

It remains to establish the inequality (26.8). By (26.9), and the triangle inequality for sums,

$$\int |f| \, dx + \varepsilon \geq \sum_{Q \in \mathcal{G}} |f(x_Q)| v(Q) \geq$$

$$\geq \left| \sum_{Q \in \mathcal{G}} f(x_Q) v(Q) \right| \geq \left| \int f \, dx \right| - \varepsilon.$$

Then since $\varepsilon$ is arbitrary, this establishes the desired inequality. This proves the corollary.

Which functions are in $R(\mathbb{R}^n)$? Begin with step functions defined below.

**Definition 26.1.3** If

$$Q \equiv \prod_{i=1}^{n} [a_i, b_i]$$

is a box, define $\text{int}(Q)$ as

$$\text{int}(Q) \equiv \prod_{i=1}^{n} (a_i, b_i).$$

$f$ is called a step function if there is a grid, $\mathcal{G}$ such that $f$ is constant on $\text{int}(Q)$ for each $Q \in \mathcal{G}$, $f$ is bounded, and $f(x) = 0$ for all $x$ outside some bounded set.

The next corollary states that step functions are in $R(\mathbb{R}^n)$ and shows the expected formula for the integral is valid.

**Corollary 26.1.4** Let $\mathcal{G}$ be a grid and let $f$ be a step function such that $f = f_Q$ on $\text{int}(Q)$ for each $Q \in \mathcal{G}$. Then $f \in R(\mathbb{R}^n)$ and

$$\int f \, dx = \sum_{Q \in \mathcal{G}} f_Q v(Q).$$

**Proof:** Let $Q$ be a box of $\mathcal{G}$,

$$Q \equiv \prod_{i=1}^{n} [a_i, b_i],$$

and suppose $g$ is a bounded function, $|g(x)| \leq C$, and $g = 0$ off $Q$, and $g = 1$ on $\text{int}(Q)$. Thus, $g$ is the simplest sort of step function. Refine $\mathcal{G}$ by including the extra points,

$$\alpha^i_{j_i} + \eta \text{ and } \alpha^i_{j_i+1} - \eta$$

for each $i = 1, \ldots, n$. Here $\eta$ is small enough that for each $i$, $\alpha^i_{j_i} + \eta < \alpha^i_{j_i+1} - \eta$. Also let $L$ denote the largest of the lengths of the sides of $Q$. Let $\mathcal{F}$ be this refined grid and denote by $Q_\eta$ the box

$$\prod_{i=1}^{n} [\alpha^i_{j_i} + \eta, \alpha^i_{j_i+1} - \eta].$$

Now define the box, $B^k$ by

$$B^k \equiv [\alpha^1_{j_1}, \alpha^1_{j_1+1}] \times \cdots \times [\alpha^{k-1}_{j_{k-1}}, \alpha^{k-1}_{j_{k-1}+1}] \times$$
In words, replace the closed interval in the \( k \)th slot used to define \( Q \) with a much thinner closed interval at one end or the other while leaving the other intervals used to define \( Q \) the same. This is illustrated in the following picture.

\[
\begin{align*}
\text{Q}_\eta & \quad \begin{array}{cccc}
\text{Q} & & & \\
& & & \\
& & & \\
& & & \\
\end{array} & \quad B_k \equiv [\alpha_{j_1}^{k-1}, \alpha_{j_1}^{k+1}] \times \cdots \times [\alpha_{j_n}^{k-1}, \alpha_{j_n}^{k+1}].
\end{align*}
\]

The important thing to notice, is that every point of \( Q \) is either in \( \text{Q}_\eta \) or one of the sets, \( B_k \). Therefore,

\[
\mathcal{L}_F (g) \geq v (Q_\eta) - \sum_{k=1}^{n} 2Cv (B_k) \geq v (Q_\eta) - 4CL^{n-1}n\eta
\]

\[
= v (Q_\eta) - K\eta
\]

(26.10)

where \( K \) is a constant which does not depend on \( \eta \).

Similarly,

\[
\mathcal{U}_F (g) \leq v (Q_\eta) + K\eta.
\]

(26.11)

This implies \( \mathcal{U}_F (g) - \mathcal{L}_F (g) < 2K\eta \) and since \( \eta \) is arbitrary, the Riemann criterion verifies that \( g \in \mathcal{R} (\mathbb{R}^n) \). Formulas (26.10) and (26.11) also verify that

\[
v (Q_\eta) \in [\mathcal{U}_F (g) - K\eta, \mathcal{L}_F (g) + K\eta] \subseteq [\mathcal{L}_F (g) - K\eta, \mathcal{U}_F (g) + K\eta].
\]

But also

\[
\int g \, dx \in [\mathcal{L}_F (g), \mathcal{U}_F (g)] \subseteq [\mathcal{L}_F (g) - K\eta, \mathcal{U}_F (g) + K\eta]
\]

and so

\[
\left| \int g \, dx - v (Q_\eta) \right| \leq 4K\eta.
\]

Now letting \( \eta \to 0 \), yields \( \int g \, dx = v (Q) \).

Now let \( f \) be as described in the statement of the Corollary. Let \( f_Q \) be the value of \( f \) on \( \text{int} (Q) \), and let \( g_Q \) be a function of the sort just considered which equals 1 on \( \text{int} (Q) \). Then \( f \) is of the form

\[
f = \sum_{Q \in \mathcal{Q}} f_Q g_Q
\]
with all but finitely many of the \( f_Q \) equal zero. Therefore, the above is really a finite sum and so by Corollary 26.1.2, \( f \in \mathcal{R}(\mathbb{R}^n) \) and

\[
\int f \, dx = \sum_{Q \in \mathcal{G}} f_Q \int g_Q \, dx = \sum_{Q \in \mathcal{G}} f_Q v(Q).
\]

There is a good deal of sloppiness inherent in the above description of a step function due to the fact that the boxes may be different but match up on an edge. It is convenient to be able to consider a more precise sort of function and this is done next.

For \( Q \) a box of the form

\[
Q = \prod_{i=1}^k [a_i, b_i],
\]

define the half open box, \( Q' \) by

\[
Q' = \prod_{i=1}^k (a_i, b_i].
\]

The reason for considering these sets is that if \( \mathcal{G} \) is a grid, the sets, \( Q' \) where \( Q \in \mathcal{G} \) are disjoint. Defining a step function, \( \phi \) as

\[
\phi(x) = \sum_{Q \in \mathcal{G}} \phi_Q \chi_{Q'}(x),
\]

the number, \( \phi_Q \) is the value of \( \phi \) on the set, \( Q' \). As before, define

\[
M_{Q'}(f) \equiv \sup \{ f(x) : x \in Q' \}, \quad m_{Q'}(f) \equiv \inf \{ f(x) : x \in Q' \}.
\]

The next lemma will be convenient a little later.

**Lemma 26.1.5** Suppose \( f \) is a bounded function which equals zero off some bounded set. Then \( f \in \mathcal{R}(\mathbb{R}^n) \) if and only if for all \( \varepsilon > 0 \) there exists a grid, \( \mathcal{G} \) such that

\[
\sum_{Q \in \mathcal{G}} (M_{Q'}(f) - m_{Q'}(f)) v(Q) < \varepsilon. \tag{26.12}
\]

**Proof:** Since \( Q' \subseteq Q \),

\[
M_{Q'}(f) - m_{Q'}(f) \leq M_Q(f) - m_Q(f)
\]

and therefore, the only if part of the equivalence is obvious.

Conversely, let \( \mathcal{G} \) be a grid such that (26.12) holds with \( \varepsilon \) replaced with \( \frac{\varepsilon}{2} \). It is necessary to show there is a grid such that (26.12) holds with no primes on the \( Q \). Let \( \mathcal{F} \) be a refinement of \( \mathcal{G} \) obtained by adding the points \( \alpha_k^i + \eta_k \) where \( \eta_k \leq \eta \) and is also chosen so small that for each \( i = 1, \ldots, n \),

\[
\alpha_k^i + \eta_k < \alpha_{k+1}^i.
\]

Then for \( Q \equiv \prod_{i=1}^n [\alpha_k^i, \alpha_{k+1}^i] \in \mathcal{G} \),

\[
\hat{Q} \equiv \prod_{i=1}^n [\alpha_k^i + \eta_k, \alpha_{k+1}^i]
\]

Let
and denote by $\hat{G}$ the collection of these smaller boxes. For each set, $Q$ in $G$ there is the smaller set, $\hat{Q}$ along with $n$ boxes, $B_k, k = 1, \ldots, n$, one of whose sides is of length $\eta_k$ and the remainder of whose sides are shorter than the diameter of $Q$ such that the set, $Q$ is the union of $\hat{Q}$ and these sets, $B_k$. Now suppose $f$ equals zero off the ball $B(0, R/2)$. Then without loss of generality, you may assume the diameter of every box in $G$ which has nonempty intersection with $B(0, R)$ is smaller than $R/3$. (If this is not so, simply refine $G$ to make it so, such a refinement leaving (26.12) valid.) Suppose there are $P$ sets of $G$ contained in $B(0, R)$ and suppose that for all $x$, $|f(x)| < C/2$.

Then

$$\sum_{Q \in F} (M_Q(f) - m_Q(f))v(Q) \leq \sum_{Q \in \hat{G}} (M_Q(f) - m_Q(f))v(Q) + \sum_{Q \in F \setminus \hat{G}} (M_Q(f) - m_Q(f))v(Q)$$

whenever $\eta$ is small enough. Since $\epsilon$ is arbitrary, $f \in \mathcal{R}(\mathbb{R}^n)$ as claimed.

**Definition 26.1.6** A bounded set, $E$, is a Jordan set in $\mathbb{R}^n$ or a contented set in $\mathbb{R}^n$ if $X_E \in \mathcal{R}(\mathbb{R}^n)$. Also, for $G$ a grid and $E$ a set, denote by $\partial_G(E)$ those boxes of $G$ which have nonempty intersection with both $E$ and $\mathbb{R}^n \setminus E$.

The next theorem is a characterization of those sets which are Jordan sets.

**Theorem 26.1.7** A bounded set, $E$, is a Jordan set if and only if for every $\epsilon > 0$ there exists a grid, $G$, such that

$$\sum_{Q \in \partial_G(E)} v(Q) < \epsilon.$$

**Proof:** If $Q \notin \partial_G(E)$, then

$$M_Q(X_E) - m_Q(X_E) = 0$$

and if $Q \in \partial_G(E)$, then

$$M_Q(X_E) - m_Q(X_E) = 1.$$

It follows that $U_G(X_E) - L_G(X_E) = \sum_{Q \in \partial_G(E)} v(Q)$ and this implies the conclusion of the theorem.

Note that if $E$ is a Jordan set and if $f \in \mathcal{R}(\mathbb{R}^n)$, then by Corollary 26.1.2, $X_Ef \in \mathcal{R}(\mathbb{R}^n)$.

**Definition 26.1.8** For $E$ a Jordan set and $fX_E \in \mathcal{R}(\mathbb{R}^n)$.

$$\int_E f \, dV \equiv \int_{\mathbb{R}^n} X_E f \, dV.$$  

A bounded set, $E$, has Jordan content 0 or content 0 if for every $\epsilon > 0$ there exists a grid, $G$ such that

$$\sum_{Q \cap E \neq \emptyset} v(Q) < \epsilon.$$

This symbol says to sum the volumes of all boxes from $G$ which have nonempty intersection with $E$.

Note that any finite union of sets having Jordan content 0 also has Jordan content 0. (Why?)
Definition 26.1.9 Let \( A \) be any subset of \( \mathbb{R}^n \). Then \( \partial A \) denotes those points, \( x \) with the property that if \( U \) is any open set containing \( x \), then \( U \) contains points of \( A \) as well as points of \( A^C \).

Corollary 26.1.10 If a bounded set, \( E \subseteq \mathbb{R}^n \) is contented, then \( \partial E \) has content 0.

Proof: Let \( \varepsilon > 0 \) be given and suppose \( E \) is contented. Then there exists a grid, \( G \) such that
\[
\sum_{Q \in \partial G(E)} v(Q) < \frac{\varepsilon}{2n + 1}. \tag{26.13}
\]

Now refine \( G \) if necessary to get a new grid, \( F \) such that all boxes from \( F \) which have nonempty intersection with \( \partial E \) have sides no larger than \( \delta \) where \( \delta \) is the smallest of all the sides of all the \( Q \) in the above sum. Recall that \( \partial G(E) \) consists of those boxes of \( G \) which have nonempty intersection with both \( E \) and \( \mathbb{R}^n \setminus E \).

Let \( x \in \partial E \). Then since the dimension is \( n \), there are at most \( 2^n \) boxes from \( F \) which contain \( x \). Furthermore, at least one of these boxes is in \( \partial F(E) \) and is therefore a subset of a box from \( \partial G(E) \). Here is why. If \( x \) is an interior point of some \( Q \in F \), then there are points of both \( E \) and \( E^C \) contained in \( Q \) and so \( x \in Q \in \partial F(E) \) and there are no other boxes from \( F \) which contain \( x \). If \( x \) is not an interior point of any \( Q \in F \), then the interior of the union of all the boxes from \( F \) which do contain \( x \) is an open set and therefore, must contain points of \( E \) and points from \( E^C \). If \( x \in E \), then one of these boxes must contain points which are not in \( E \) since otherwise, \( x \) would fail to be in \( \partial E \). Pick that box. It is in \( \partial F(E) \) and contains \( x \). On the other hand, if \( x \notin E \), one of these boxes must contain points of \( E \) since otherwise, \( x \) would fail to be in \( \partial E \). Pick that box. This shows that every set from \( F \) which contains a point of \( \partial E \) shares this point with a box of \( \partial G(E) \). Let the boxes from \( \partial G(E) \) be \( \{P_1, \ldots, P_m \} \). Let \( S(P_i) \) denote those sets of \( F \) which contain a point of \( \partial E \) in common with \( P_i \). Then if \( Q \in S(P_i) \), either \( Q \subseteq P_i \) or it intersects \( P_i \) on one of its \( 2n \) faces. Therefore, the sum of the volumes of those boxes of \( S(P_i) \) which intersect \( P_i \) on a particular face of \( P_i \) is no larger than \( v(P_i) \). Consequently,
\[
\sum_{Q \in S(P_i)} v(Q) \leq 2n v(P_i) + v(P_i)
\]
and so for \( Q \in F \),
\[
\sum_{Q \cap \partial E \neq \emptyset} v(Q) = \sum_{i=1}^{m} \sum_{Q \in S(P_i)} v(Q) \leq \sum_{i=1}^{m} (2n + 1) v(P_i) < \varepsilon
\]
from (26.13). This proves the corollary.

Theorem 26.1.11 If a bounded set, \( E \), has Jordan content 0, then \( E \) is a Jordan set and if \( f \) is any bounded function defined on \( E \), then \( f\chi_E \in \mathcal{R}(\mathbb{R}^n) \) and
\[
\int_E f \, d\mathcal{V} = 0.
\]

Proof: Let \( \varepsilon > 0 \). Then let \( G \) be a grid such that
\[
\sum_{Q \cap E \neq \emptyset} v(Q) < \varepsilon.
\]
Then every set of \( \partial G(E) \) contains a point of \( E \) so
\[
\sum_{Q \in \partial G(E)} v(Q) \leq \sum_{Q \cap E \neq \emptyset} v(Q) < \varepsilon
\]
and since $\varepsilon$ was arbitrary, this shows from Theorem 26.1.7 that $E$ is a Jordan set. Now let $M$ be a positive number larger than all values of $f$, let $m$ be a negative number smaller than all values of $f$ and let $\varepsilon > 0$ be given. Let $\mathcal{G}$ be a grid with

$$\sum_{Q \cap E \neq \emptyset} v(Q) < \frac{\varepsilon}{1 + (M - m)}.$$  

Then

$$U_{\mathcal{G}}(f_{\chi_E}) \leq \sum_{Q \cap E \neq \emptyset} Mv(Q) \leq \frac{\varepsilon M}{1 + (M - m)}$$

and

$$L_{\mathcal{G}}(f_{\chi_E}) \geq \sum_{Q \cap E \neq \emptyset} mv(Q) \geq \frac{\varepsilon m}{1 + (M - m)}$$

and so

$$U_{\mathcal{G}}(f_{\chi_E}) - L_{\mathcal{G}}(f_{\chi_E}) \leq \sum_{Q \cap E \neq \emptyset} Mv(Q) - \sum_{Q \cap E \neq \emptyset} mv(Q)$$

$$= (M - m) \sum_{Q \cap E \neq \emptyset} v(Q) < \frac{\varepsilon (m - N)}{1 + (M - m)} < \varepsilon.$$  

This shows $f_{\chi_E} \in \mathcal{R}(\mathbb{R}^n)$. Now also,

$$m \varepsilon \leq \int f_{\chi_E} dV \leq M \varepsilon$$

and since $\varepsilon$ is arbitrary, this shows

$$\int_E f \, dV \equiv \int f_{\chi_E} dV = 0$$

and proves the theorem.

**Corollary 26.1.12** If $f_{\chi_{E_i}} \in \mathcal{R}(\mathbb{R}^n)$ for $i = 1, 2, \ldots, r$ and for all $i \neq j, E_i \cap E_j$ is either the empty set or a set of Jordan content 0, then letting $F \equiv \bigcup_{i=1}^{r} E_i$, it follows $f_{\chi_F} \in \mathcal{R}(\mathbb{R}^n)$ and

$$\int f_{\chi_F} dV \equiv \int f \, dV = \sum_{i=1}^{r} \int_{E_i} f \, dV.$$  

**Proof:** By Corollary 26.1.2, this is true if $r = 1$. Suppose it is true for $r$. It will be shown that it is true for $r + 1$. Let $F_r = \bigcup_{i=1}^{r} E_i$ and let $F_{r+1}$ be defined similarly. By the induction hypothesis, $f_{\chi_{F_r}} \in \mathcal{R}(\mathbb{R}^n)$. Also, since $F_r$ is a finite union of the $E_i$, it follows that $F_r \cap E_{r+1}$ is either empty or a set of Jordan content 0.

$$-f_{\chi_{F_r \cap E_{r+1}}} + f_{\chi_{F_r}} + f_{\chi_{E_{r+1}}} = f_{\chi_{F_{r+1}}}$$

and by Theorem 26.1.11 each function on the left is in $\mathcal{R}(\mathbb{R}^n)$ and the first one on the left has integral equal to zero. Therefore,

$$\int f_{\chi_{E_{r+1}}} dV = \int f_{\chi_{F_r}} dV + \int f_{\chi_{E_{r+1}}} dV$$

which by induction equals

$$\sum_{i=1}^{r} \int_{E_i} f \, dV + \int_{E_{r+1}} f \, dV = \sum_{i=1}^{r+1} \int_{E_i} f \, dV.$$  

and this proves the corollary.

What functions in addition to step functions are integrable? As in the case of integrals of functions of one variable, this is an important question. It turns out that the Riemann integrable functions are characterized by being continuous except on a very small set. To begin with it is necessary to define the oscillation of a function.

**Definition 26.1.13** Let \( f \) be a function defined on \( \mathbb{R}^n \) and let

\[
\omega_{f,r}(x) \equiv \sup \{|f(z) - f(y)| : z, y \in B(x,r)\}.
\]

This is called the oscillation of \( f \) on \( B(x,r) \). Note that this function of \( r \) is decreasing in \( r \). Define the oscillation of \( f \) as

\[
\omega_f(x) \equiv \lim_{r \to 0^+} \omega_{f,r}(x).
\]

Note that as \( r \) decreases, the function, \( \omega_{f,r}(x) \) decreases. It is also bounded below by 0 and so the limit must exist and equals \( \inf \{\omega_{f,r}(x) : r > 0\} \). (Why?) Then the following simple lemma whose proof follows directly from the definition of continuity gives the reason for this definition.

**Lemma 26.1.14** A function, \( f \), is continuous at \( x \) if and only if \( \omega_f(x) = 0 \).

This concept of oscillation gives a way to define how discontinuous a function is at a point. The discussion will depend on the following fundamental lemma which gives the existence of something called the Lebesgue number.

**Definition 26.1.15** Let \( \mathcal{C} \) be a set whose elements are sets of \( \mathbb{R}^n \) and let \( K \subseteq \mathbb{R}^n \). The set, \( \mathcal{C} \) is called a cover of \( K \) if every point of \( K \) is contained in some set of \( \mathcal{C} \). If the elements of \( \mathcal{C} \) are open sets, it is called an open cover.

**Lemma 26.1.16** Let \( K \) be sequentially compact and let \( \mathcal{C} \) be an open cover of \( K \). Then there exists \( r > 0 \) such that whenever \( x \in K \), \( B(x,r) \) is contained in some set of \( \mathcal{C} \).

**Proof:** Suppose this is not so. Then letting \( r_n = 1/n \), there exists \( x_n \in K \) such that \( B(x_n,r_n) \) is not contained in any set of \( \mathcal{C} \). Since \( K \) is sequentially compact, there is a subsequence, \( x_{n_k} \) which converges to a point, \( x \in K \). But there exists \( \delta > 0 \) such that \( B(x,\delta) \subseteq U \) for some \( U \in \mathcal{C} \). Let \( k \) be so large that \( 1/k < \delta/2 \) and \( |x_{n_k} - x| < \delta/2 \) also. Then if \( z \in B(x_{n_k},r_{n_k}) \), it follows

\[
|z - x| \leq |z - x_{n_k}| + |x_{n_k} - x| < \frac{\delta}{2} + \frac{\delta}{2} = \delta
\]

and so \( B(x_{n_k},r_{n_k}) \subseteq U \) contrary to supposition. Therefore, the desired number exists after all.

**Theorem 26.1.17** Let \( f \) be a bounded function which equals zero off a bounded set and let \( W \) denote the set of points where \( f \) fails to be continuous. Then \( f \in \mathcal{R}(\mathbb{R}^n) \) if \( W \) has content zero. That is, for all \( \varepsilon > 0 \) there exists a grid, \( \mathcal{G} \) such that

\[
\sum_{Q \in \mathcal{G}_W} v(Q) < \varepsilon
\]

where

\[
\mathcal{G}_W \equiv \{Q \in \mathcal{G} : Q \cap W \neq \emptyset\}.
\]
**Proof:** Let \(|f(x)| < C/2\) for all \(x \in \mathbb{R}^n\), let \(\varepsilon > 0\) be given, and let \(G\) be a grid which satisfies (26.14). Since \(f\) equals zero off some bounded set, there exists \(R\) such that \(f\) equals zero off of \(B\left(0, \frac{R}{2}\right)\). Thus \(W \subseteq B\left(0, \frac{R}{2}\right)\). Also note that if \(G\) is a grid for which (26.14) holds, then this inequality continues to hold if \(G\) is replaced with a refined grid. Therefore, you may assume the diameter of every box in \(G\) which intersects \(B(0,R)\) is less than \(\frac{R}{3}\). Since \(W\) is bounded, \(G_W\) contains only finitely many boxes. Letting

\[
Q = \prod_{i=1}^{n} [a_i, b_i]
\]

be one of these boxes, enlarge the box slightly as indicated in the following picture.

![Diagram](image)

The enlarged box is an open set of the form,

\[
\tilde{Q} = \prod_{i=1}^{n} (a_i - \eta_i, b_i + \eta_i)
\]

where \(\eta_i\) is chosen small enough that if

\[
\prod_{i=1}^{n} (b_i + \eta_i - (a_i - \eta_i)) \equiv v(\tilde{Q}),
\]

then

\[
\sum_{Q \in G_W} v(\tilde{Q}) < \varepsilon.
\]

For each \(x \in \mathbb{R}^n\), let \(r_x\) be such that

\[
\omega_{f,r_x}(x) < \varepsilon + \omega_f(x). \tag{26.15}
\]

Now let \(\mathcal{C}\) denote all intersections of the form \(Q \cap B(x, r_x)\) such that \(x \in B(0, R)\) so that \(\mathcal{C}\) is an open cover of the compact set, \(B(0, R)\). Let \(\delta\) be a Lebesgue number for this open cover of \(B(0, R)\) and let \(F\) be a refinement of \(G\) such that every box in \(F\) has diameter less than \(\delta\).

Now let \(F_1\) consist of those boxes of \(F\) which have nonempty intersection with \(B(0, R/2)\). Thus all boxes of \(F_1\) are contained in \(B(0, R)\) and each one is contained in some set of \(\mathcal{C}\).

Now let \(\mathcal{C}_W\) be those open sets of \(\mathcal{C}\), \(Q \cap B(x, r_x)\), for which \(x \in W\) and let \(F_W\) be those sets of \(F_1\) which are subsets of some set of \(\mathcal{C}_W\). Then

\[
U_F(f) - L_F(f) = \sum_{Q \in F_W} (M_Q(f) - m_Q(f)) v(Q) + \sum_{Q \in F_1 \setminus F_W} (M_Q(f) - m_Q(f)) v(Q).
\]

If \(Q \in F_1 \setminus F_W\), then \(Q\) must be a subset of some set of \(\mathcal{C} \setminus \mathcal{C}_W\) since it is not in any set of \(\mathcal{C}_W\). Therefore, from (26.15) and the observation that \(x \notin W\),

\[
M_Q(f) - m_Q(f) \leq \varepsilon.
\]
Suppose some a be the case that for all \( i \),

\[
U_{\mathcal{G}}(f) - L_{\mathcal{G}}(f) \leq \sum_{Q \in \mathcal{F}_W} C\nu(Q) + \sum_{Q \in \mathcal{F}_1 \setminus \mathcal{F}_W} \varepsilon \nu(Q)
\]

\[
\leq C\varepsilon + \varepsilon (2R)^n.
\]

Since \( \varepsilon \) is arbitrary, this proves the theorem.\(^1\)

From Theorem 26.1.7 you get a pretty good idea of what constitutes a contented set. These sets are essentially those which have thin boundaries. Most sets you are likely to think of will fall in this category. However, it is good to give specific examples of sets which are contented.

**Theorem 26.1.18** Suppose \( E \) is a bounded contented set in \( \mathbb{R}^n \) and \( f, g : E \to \mathbb{R} \) are two functions satisfying \( f(x) \geq g(x) \) for all \( x \in E \) and \( fX_E \) and \( gX_E \) are both in \( \mathcal{R}(\mathbb{R}^n) \). Now define

\[
P \equiv \{(x, x_{n+1}) : x \in E \text{ and } g(x) \leq x_{n+1} \leq f(x)\}.
\]

Then \( P \) is a contented set in \( \mathbb{R}^{n+1} \).

**Proof:** Let \( \mathcal{G} \) be a grid such that for \( k = f, g \),

\[
U_{\mathcal{G}}(k) - L_{\mathcal{G}}(k) < \varepsilon/2.
\]

(26.16)

Let the boxes of \( \mathcal{G} \) which have nonempty intersection with \( E \) be \( \{Q_1, \cdots, Q_m\} \) and let \( \{a_i\}_{i=-\infty}^{\infty} \) be a sequence on \( \mathbb{R} \), \( a_i < a_{i+1} \) for all \( i \), which includes

\[
M_{Q_j}(fX_E), M_{Q_j}(gX_E), m_{Q_j}(fX_E), m_{Q_j}(gX_E)
\]

for all \( j = 1, \cdots, m \). Now define a grid on \( \mathbb{R}^{n+1} \) as follows.

\[
\mathcal{G}' \equiv \{Q \times [a_i, a_{i+1}] : Q \in \mathcal{G}, i \in \mathbb{Z}\}
\]

In words, this grid consists of all possible boxes of the form \( Q \times [a_i, a_{i+1}] \) where \( Q \in \mathcal{G} \) and \( a_i \) is a term of the sequence just described. It is necessary to verify that \( \mathcal{X}_P \in \mathcal{R}(\mathbb{R}^{n+1}) \).

This is done by showing that \( U_{\mathcal{G}'}(\mathcal{X}_P) - L_{\mathcal{G}'}(\mathcal{X}_P) < \varepsilon \) and then noting that \( \varepsilon > 0 \) was arbitrary. For \( \mathcal{G}' \) just described, denote by \( Q' \) a box in \( \mathcal{G}' \). Thus \( Q' = Q \times [a_i, a_{i+1}] \) for some \( i \).

\[
U_{\mathcal{G}'}(\mathcal{X}_P) - L_{\mathcal{G}'}(\mathcal{X}_P) = \sum_{Q' \in \mathcal{G}'} (M_{Q'}(\mathcal{X}_P) - m_{Q'}(\mathcal{X}_P)) v_{n+1}(Q')
\]

\[
= \sum_{i=-\infty}^{\infty} \sum_{j=1}^{m} (M_{Q_j}(\mathcal{X}_P) - m_{Q_j}(\mathcal{X}_P)) v_n(Q_j)(a_{i+1} - a_i)
\]

and all sums are bounded because the functions, \( f \) and \( g \) are given to be bounded. Therefore, there are no limit considerations needed here. Thus

\[
U_{\mathcal{G}'}(\mathcal{X}_P) - L_{\mathcal{G}'}(\mathcal{X}_P) = \sum_{j=1}^{m} \sum_{i=-\infty}^{\infty} (M_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P))(a_{i+1} - a_i).
\]

\(^1\) In fact one cannot do any better. It can be shown that if a function is Riemann integrable, then it must be the case that for all \( \varepsilon > 0 \), (26.14) is satisfied for some grid, \( \mathcal{G} \). This along with what was just shown is known as Lebesgue’s theorem after Lebesgue who discovered it in the early years of the twentieth century. Actually, he also invented a far superior integral which has been the integral of serious mathematicians since that time.
Consider the inside sum with the aid of the following picture.

In this picture, the little rectangles represent the boxes $Q_j \times [a_i, a_{i+1}]$ for fixed $j$. The part of $P$ having $x$ contained in $Q_j$ is between the two surfaces, $x_{n+1} = g(x)$ and $x_{n+1} = f(x)$ and there is a zero placed in those boxes for which $M_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) = 0$. You see, $\mathcal{X}_P$ has either the value of 1 or the value of 0 depending on whether $(x, y)$ is contained in $P$. For the boxes shown with 0 in them, either all of the box is contained in $P$ or none of the box is contained in $P$. Either way, $M_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) = 0$ on these boxes. However, on the boxes intersected by the surfaces, the value of $M_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P)$ is 1 because there are points in this box which are not in $P$ as well as points which are in $P$. Because of the construction of $G'$ which included all values of $M_{Q_j}(f \mathcal{X}_E), M_{Q_j}(g \mathcal{X}_E), m_{Q_j}(f \mathcal{X}_E), m_{Q_j}(g \mathcal{X}_E)$ for all $j = 1, \ldots, m$,

$$
\sum_{i = -\infty}^{\infty} \left( M_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) \right) (a_{i+1} - a_i) \leq \sum_{\{i : m_{Q_j}(g) \leq a_i < M_{Q_j}(g)\}} 1 (a_{i+1} - a_i) + \sum_{\{i : m_{Q_j}(f) \leq a_i < M_{Q_j}(f)\}} 1 (a_{i+1} - a_i) = (M_{Q_j}(g) - m_{Q_j}(g)) + (M_{Q_j}(f) - m_{Q_j}(f)).
$$

(Note the inequality.) Therefore, by (26.16),

$$
U_{G'}(\mathcal{X}_P) - \mathcal{L}_{G'}(\mathcal{X}_P) \leq \sum_{j = 1}^{m} v_n(Q_j) \left[ (M_{Q_j}(g) - m_{Q_j}(g)) + (M_{Q_j}(f) - m_{Q_j}(f)) \right] = U_G(f) - \mathcal{L}_G(f) + U_G(g) - \mathcal{L}_G(g) < \varepsilon/2 + \varepsilon/2 = \varepsilon.
$$

Since $\varepsilon > 0$ is arbitrary, this proves the theorem.

**Corollary 26.1.19** Suppose $f$ and $g$ are continuous functions defined on $E$, a contented set in $\mathbb{R}^n$ and that $g(x) \leq f(x)$ for all $x \in E$. Then

$$
P \equiv \{ (x, x_{n+1}) : x \in E \text{ and } g(x) \leq x_{n+1} \leq f(x) \}
$$

is a contented set in $\mathbb{R}^n$.

**Proof:** Extend $f$ and $g$ to equal 0 off $E$. The set of discontinuities of $f$ and $g$ is contained in $\partial E$ and Corollary 26.1.10 on Page 486 implies this is a set of content 0. Therefore, from Theorem 26.1.17, for $k = f, g$, it follows that $k\mathcal{X}_E$ is in $\mathcal{R}(\mathbb{R}^n)$ because the set of discontinuities is contained in $\partial E$. The conclusion now follows from Theorem 26.1.18. This proves the corollary.
As an example of how this can be applied, it is obvious a closed interval is a contented set in \( \mathbb{R} \). Therefore, if \( f, g \) are two continuous functions with \( f(x) \geq g(x) \) for \( x \in [a, b] \), it follows from the above theorem or its corollary that the set,

\[
P_1 \equiv \{(x, y) : g(x) \leq y \leq f(x)\}
\]

is a contented set in \( \mathbb{R}^2 \). Now using the theorem and corollary again, suppose \( f_1(x, y) \geq g_1(x, y) \) for \((x, y) \in P_1\) and \( f, g \) are continuous. Then the set

\[
P_2 \equiv \{(x, y, z) : g_1(x, y) \leq z \leq f_1(x, y)\}
\]

is a contented set in \( \mathbb{R}^3 \). Clearly you can continue this way obtaining examples of contented sets.

Note that as a special case of Corollary 26.1.4 on Page 482, it follows that every box is a contented set.

### 26.2 Iterated Integrals

To evaluate an \( n \) dimensional Riemann integral, one uses iterated integrals. Formally, an iterated integral is defined as follows. For \( f \) a function defined on \( \mathbb{R}^{n+m} \),

\[
y \to f(x, y)
\]

is a function of \( y \) for each \( x \in \mathbb{R}^{n+m} \). Therefore, it might be possible to integrate this function of \( y \) and write

\[
\int_{\mathbb{R}^m} f(x, y) \, dV_y.
\]

Now the result is clearly a function of \( x \) and so, it might be possible to integrate this and write

\[
\int_{\mathbb{R}^n} \int_{\mathbb{R}^m} f(x, y) \, dV_y \, dV_x.
\]

This symbol is called an iterated integral, because it involves the iteration of two lower dimensional integrations. Under what conditions are the two iterated integrals equal to the integral

\[
\int_{\mathbb{R}^{n+m}} f(z) \, dV?
\]

**Definition 26.2.1** Let \( \mathcal{G} \) be a grid on \( \mathbb{R}^{n+m} \) defined by the \( n + m \) sequences,

\[
\{\alpha_i^k\}_{k=-\infty}^{\infty} : i = 1, \ldots, n + m.
\]

Let \( \mathcal{G}_n \) be the grid on \( \mathbb{R}^n \) obtained by considering only the first \( n \) of these sequences and let \( \mathcal{G}_m \) be the grid on \( \mathbb{R}^m \) obtained by considering only the last \( m \) of the sequences. Thus a typical box in \( \mathcal{G}_m \) would be

\[
\prod_{i=n+1}^{n+m} [\alpha_i^k, \alpha_i^{k+1}], \ k_i \geq n + 1
\]

and a box in \( \mathcal{G}_n \) would be of the form

\[
\prod_{i=1}^{n} [\alpha_i^k, \alpha_i^{k+1}], \ k_i \leq n.
\]
Lemma 26.2.2 Let $G_n$ and $G_m$ be the grids defined above. Then

$$G = \{ R \times P : R \in G_n \text{ and } P \in G_m \}.$$  

**Proof:** If $Q \in G$, then $Q$ is clearly of this form. On the other hand, if $R \times P$ is one of the sets described above, then from the above description of $R$ and $P$, it follows $R \times P$ is one of the sets of $G$.

Now let $G$ be a grid on $\mathbb{R}^{n+m}$ and suppose

$$\phi(z) = \sum_{Q \in G} \phi_Q X_Q (z) \quad (26.17)$$

where $\phi_Q$ equals zero for all but finitely many $Q$. Thus $\phi$ is a step function. Recall that for

$$Q = \prod_{i=1}^{n+m} [a_i, b_i], \quad Q' = \prod_{i=1}^{n+m} (a_i, b_i)$$

Letting $(x, y) = z$, Lemma 26.2.2 implies

$$\phi(z) = \phi(x, y) = \sum_{R \in G_n} \sum_{P \in G_m} \phi_{R \times P} X_R (x) \mathbb{1}_P (y)$$

$$= \sum_{R \in G_n} \sum_{P \in G_m} \phi_{R \times P} X_R (x) \mathbb{1}_P (y). \quad (26.18)$$

For a function of two variables, $h$, denote by $h(\cdot, y)$ the function, $x \mapsto h(x, y)$ and $h(x, \cdot)$ the function $y \mapsto h(x, y)$. The following lemma is a preliminary version of Fubini’s theorem.

**Lemma 26.2.3** Let $\phi$ be a step function as described in (26.17). Then

$$\phi(x, \cdot) \in \mathcal{R} (\mathbb{R}^m), \quad (26.19)$$

$$\int_{\mathbb{R}^n} \phi(\cdot, y) \, dV_y \in \mathcal{R} (\mathbb{R}^n), \quad (26.20)$$

and

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}^m} \phi(x, y) \, dV_y \, dV_x = \int_{\mathbb{R}^{n+m}} \phi(z) \, dV. \quad (26.21)$$

**Proof:** To verify (26.19), note that $\phi(x, \cdot)$ is the step function

$$\phi(x, \cdot) = \sum_{P \in G_m} \phi_{R \times P} \mathbb{1}_P (y).$$

Where $x \in \mathbb{R}$. By Corollary 26.1.4, this verifies (26.19). From the description in (26.18) and this corollary,

$$\int_{\mathbb{R}^n} \phi(x, y) \, dV_y = \sum_{R \in G_n} \sum_{P \in G_m} \phi_{R \times P} X_R (x) v(P)$$

$$= \sum_{R \in G_n} \left( \sum_{P \in G_m} \phi_{R \times P} v(P) \right) X_R (x), \quad (26.22)$$
another step function. Therefore, Corollary 26.1.4 applies again to verify (26.20). Finally, (26.22) implies
\[
\int_{\mathbb{R}^n} \int_{\mathbb{R}^m} \phi(x, y) \, dV_y \, dV_x = \sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} \phi_{R \times P} (P) \, v(R)
\]
\[
= \sum_{Q \in \mathcal{G}} \phi_Q (Q) = \int_{\mathbb{R}^{n+m}} \phi (z) \, dV.
\]
and this proves the lemma.

From (26.22),
\[
M_{R_1'} \left( \int_{\mathbb{R}^m} \phi (\cdot, y) \, dV_y \right) \equiv \sup \left\{ \sum_{R \in \mathcal{G}_n} \left( \sum_{P \in \mathcal{G}_m} \phi_{R \times P} (P) \right) X_R (x) : x \in R_1' \right\}
\]
\[
= \sum_{P \in \mathcal{G}_m} \phi_{R_1 \times P} (P).
\]
(26.23)

Similarly,
\[
m_{R_1'} \left( \int_{\mathbb{R}^m} \phi (\cdot, y) \, dV_y \right) \equiv \inf \left\{ \sum_{R \in \mathcal{G}_n} \left( \sum_{P \in \mathcal{G}_m} \phi_{R \times P} (P) \right) X_R (x) : x \in R_1' \right\}
\]
\[
= \sum_{P \in \mathcal{G}_m} \phi_{R_1 \times P} (P).
\]
(26.24)

**Theorem 26.2.4 (Fubini)** Let \( f \in \mathcal{C}(\mathbb{R}^{n+m}) \) and suppose also that \( f(x, \cdot) \in \mathcal{C}(\mathbb{R}^m) \) for each \( x \). Then
\[
\int_{\mathbb{R}^m} f (x, \cdot) \, dV_y \in \mathcal{C}(\mathbb{R}^n)
\]
(26.25)
and
\[
\int_{\mathbb{R}^{n+m}} f (z) \, dV = \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} f (x, y) \, dV_y \, dV_x.
\]
(26.26)

**Proof:** Let \( \mathcal{G} \) be a grid such that \( \mathcal{U}_\mathcal{G} (f) - \mathcal{L}_\mathcal{G} (f) < \varepsilon \) and let \( \mathcal{G}_n \) and \( \mathcal{G}_m \) be as defined above. Let
\[
\phi (z) \equiv \sum_{Q \in \mathcal{G}} M_Q (f) \chi_Q (z), \quad \psi (z) \equiv \sum_{Q \in \mathcal{G}} m_Q (f) \chi_Q (z).
\]
By Corollary 26.1.4, and the observation that \( M_Q (f) \leq M_Q (f) \) and \( m_Q (f) \geq m_Q (f) \),
\[
\mathcal{U}_\mathcal{G} (f) \geq \int \phi \, dV, \quad \mathcal{L}_\mathcal{G} (f) \leq \int \psi \, dV.
\]
Also \( f (z) \in (\psi (z), \phi (z)) \) for all \( z \). Thus from (26.23),
\[
M_{R'} \left( \int_{\mathbb{R}^m} f (x, y) \, dV_y \right) \leq M_{R'} \left( \int_{\mathbb{R}^m} \phi (x, y) \, dV_y \right) = \sum_{P \in \mathcal{G}_m} M_{R' \times P} (f) \, v(P)
\]
and from (26.24),
\[
m_{R'} \left( \int_{\mathbb{R}^m} f (x, y) \, dV_y \right) \geq m_{R'} \left( \int_{\mathbb{R}^m} \psi (x, y) \, dV_y \right) = \sum_{P \in \mathcal{G}_m} m_{R' \times P} (f) \, v(P).
\]
Therefore,
\[
\sum_{R \in \mathcal{G}_n} \left[ M_{R'} \left( \int_{R^n} f(x,y) \, dy \right) - m_{R'} \left( \int_{R^n} f(x,y) \, dy \right) \right] v(R) \leq \\
\sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} \left[ M_{R \times P'} (f) - m_{R \times P'} (f) \right] v(P) v(R) \leq U_G (f) - L_G (f) < \varepsilon.
\]

This shows, from Lemma 26.1.5, that \( \int_{R^n} f(x,y) \, dy \in R (R^n) \). It remains to verify (26.26).

First note
\[
\int_{R^{n+1}} f(z) \, dz \in [L_G (f), U_G (f)].
\]

Next, by Lemma 26.2.3,
\[
L_G (f) \leq \int_{R^{n+1}} \psi \, dV = \int_{R^n} \int_{R^n} \psi \, dy \, dx \leq \int_{R^n} \int_{R^n} f(x,y) \, dy \, dx \\
\leq \int_{R^n} \int_{R^n} \phi (x,y) \, dy \, dx = \int_{R^{n+1}} \phi \, dV \leq U_G (f).
\]

Therefore,
\[
\left| \int_{R^n} \int_{R^n} f(x,y) \, dy \, dx - \int_{R^{n+1}} f(z) \, dz \right| \leq \varepsilon
\]
and since \( \varepsilon > 0 \) is arbitrary, this proves Fubini’s theorem\(^2\).

**Corollary 26.2.5** Suppose \( E \) is a bounded contented set in \( R^n \) and let \( \phi, \psi \) be continuous functions defined on \( E \) such that \( \phi (x) \geq \psi (x) \). Also suppose \( f \) is a continuous bounded function defined on the set,
\[
P \equiv \{ (x,y) : \psi (x) \leq y \leq \phi (x) \},
\]
It follows \( f(P) \in R (R^{n+1}) \) and
\[
\int_{P} f \, dV = \int_{E} \int_{\psi (x)}^{\phi (x)} f(x,y) \, dy \, dx.
\]

**Proof:** Since \( f \) is continuous, there is no problem in writing \( f(x,\cdot) \in R (R^1) \). Also, \( f(P) \in R (R^{n+1}) \) because \( P \) is contented thanks to Corollary 26.1.19. Therefore, by Fubini’s theorem
\[
\int_{P} f \, dV = \int_{R^n} f(P) \, dy \, dx \\
= \int_{E} \int_{\psi (x)}^{\phi (x)} f(x,y) \, dy \, dx
\]
proving the corollary.

Other versions of this corollary are immediate and should be obvious whenever encountered.

\(^2\)Actually, Fubini’s theorem usually refers to a much more profound result in the theory of Lebesgue integration.
26.2.1 Some Observations

Some of the above material is very technical. This is because it gives complete answers to the fundamental questions on existence of the integral and related theoretical considerations. It was realized early in the twentieth century that these difficulties occur because from the point of view of mathematics, this is not the right way to define an integral! Better results are obtained much more easily using the Lebesgue integral. Many of the technicalities related to Jordan content disappear almost magically when the right integral is used. However, the Lebesgue integral is much more abstract than the Riemann integral and it is not traditional to consider it in a beginning calculus course. If you are interested in the fundamental properties of the integral and the theory behind it, you should abandon the Riemann integral which is an antiquated relic and begin to study the integral of the last century. One of the best introductions to it is in [22]. Another very good source is [11]. This advanced calculus text does everything in terms of the Lebesgue integral and never bothers to struggle with the inferior Riemann integral. A more general treatment is found in [17], [18], [23], and [19]. There is also a still more general integral called the generalized Riemann integral. A recent book on this subject is [5].
Part III

Further Topics In Linear Algebra
Rank Of A Matrix

27.0.2 Outcomes

1. Recognize and find the row reduced echelon form of a matrix.
2. Determine the rank of a matrix.
3. Characterize the solution set to a matrix using rank
4. Argue that a homogeneous linear system always has a solution and find the solutions.

27.1 The Row Reduced Echelon Form Of A Matrix

Recall the row operations used to solve a system of equations which were presented earlier.

Definition 27.1.1 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 itself.

These row operations were first used to solve systems of equations. Next they were used to find the inverse of a matrix and this was closely related to the first application. Finally they were shown to be the only practical way to evaluate large determinants.

There is a particularly simple form which you try to reach when doing row operations. It is called the row reduced echelon form. It involves doing row operations till the matrix is in reduced echelon form and then multiplying each nonzero row by the multiplicative inverse of the leading entry.

Definition 27.1.2 A matrix is in row reduced echelon form if it is in reduced echelon form and in addition, the leading entry in each nonzero row is 1.

It turns out some remarkable things are true about the row reduced echelon form, the main one being that it is unique. That is, every matrix can be row reduced to a unique row reduced echelon form. Of course this is not true of the echelon form or the reduced echelon form. The significance of this is that it becomes possible to use the definite article in referring to the row reduced echelon form and hence important conclusions about the original matrix may be logically deduced from an examination of its unique row reduced echelon form. Another way to say the same thing as expressed in the above definition is the following which turns out to be an easier description when proving the major assertions about the row reduced echelon form.
Definition 27.1.3 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, $e_1, e_2, \ldots, e_k$.

Theorem 27.1.4 Let $A$ be an $m \times n$ matrix. Then $A$ has a row reduced echelon form determined by a simple process.

Proof: Viewing the columns of $A$ from left to right take the first nonzero column. Pick a nonzero entry in this column and switch the row containing this entry with the top row of $A$. Now divide this new top row by the value of this nonzero entry to get a 1 in this position and then use row operations to make all entries below this element equal to zero. Thus the first nonzero column is now $e_1$. Denote the resulting matrix by $A_1$. Consider the submatrix of $A_1$ to the right of this column and below the first row. Do exactly the same thing for it that was done for $A$. This time the $e_1$ will refer to $F^{m-1}$. Use this 1 and row operations to zero out every element above it in the rows of $A_1$. Call the resulting matrix, $A_2$. Thus $A_2$ satisfies the conditions of the above definition up to the column just encountered. Continue this way till every column has been dealt with and the result must be in row reduced echelon form.

The following diagram illustrates the above procedure. Say the matrix looked something like the following.

\[
\begin{pmatrix}
0 & * & * & * & * & * \\
0 & * & * & * & * & * \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & * & * & * & * & *
\end{pmatrix}
\]

First step would yield something like

\[
\begin{pmatrix}
0 & 1 & * & * & * & * \\
0 & 0 & * & * & * & * \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & * & * & * & *
\end{pmatrix}
\]

For the second step you look at the lower right corner as described,

\[
\begin{pmatrix}
* & * & * & * & * \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
* & * & * & * & *
\end{pmatrix}
\]

and if the first column consists of all zeros but the next one is not all zeros, you would get something like this.

\[
\begin{pmatrix}
0 & 1 & * & * \\
\vdots & \vdots & \vdots & \vdots \\
0 & 0 & * & *
\end{pmatrix}
\]

Thus, after zeroing out the term in the top row above the 1, you get the following for the next step in the computation of the row reduced echelon form for the original matrix.

\[
\begin{pmatrix}
0 & 1 & 0 & * & * \\
0 & 0 & 1 & * & * \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & *
\end{pmatrix}
\]
Next you look at the lower right matrix below the top two rows and to the right of the first four columns and repeat the process.

Recall the following definition which was discussed earlier.

**Definition 27.1.5** The first pivot column of $A$ is the first nonzero column of $A$. The next pivot column is the first column after this which becomes $e_2$ in the row reduced echelon form. The third is the next column which becomes $e_3$ in the row reduced echelon form and so forth.

There are three choices for row operations at each step in the above theorem. A natural question is whether the same row reduced echelon matrix always results in the end from following the above algorithm applied in any way. The next corollary says this is the case.

In rough terms, the following lemma states that linear relationships between columns in a matrix are preserved by row operations.

**Lemma 27.1.6** Let $A$ and $B$ be two $m \times n$ matrices and suppose $B$ results from a row operation applied to $A$. Then the $k^{th}$ column of $B$ is a linear combination of the $i_1, \ldots, i_r$ columns of $B$ if and only if the $k^{th}$ column of $A$ is a linear combination of the $i_1, \ldots, i_r$ columns of $A$. Furthermore, the scalars in the linear combination are the same. (The linear relationship between the $k^{th}$ column of $A$ and the $i_1, \ldots, i_r$ columns of $A$ is the same as the linear relationship between the $k^{th}$ column of $B$ and the $i_1, \ldots, i_r$ columns of $B$.)

**Proof:** This is obvious in the case of the first two row operations and a little less obvious in the case of the third. Therefore, consider the third. Suppose the $s^{th}$ row of $B$ equals the $s^{th}$ row of $A$ added to $c$ times the $q^{th}$ row of $A$. Therefore,

$$B_{ij} = A_{ij} \text{ if } i \neq s, B_{sj} = A_{sj} + cA_{qj}.$$  

The assumption about the $k^{th}$ column of $B$ is equivalent to saying that for each $p$,

$$B_{pk} = \sum_{j=1}^{r} \alpha_j B_{pj}.$$  

(27.1)

For $p \neq s$, this is equivalent to saying

$$A_{pk} = \sum_{j=1}^{r} \alpha_j A_{pj}.$$  

(27.2)

because for these values of $p$, $B_{pj} = A_{pj}$. For $p = s$, this is equivalent to saying

$$A_{sk} + cA_{qk} = \sum_{j=1}^{r} \alpha_j (A_{sij} + cA_{qij}).$$  

(27.3)

but from (27.2), applied to $p = q$,

$$cA_{qk} = c \sum_{j=1}^{r} \alpha_j A_{qij}$$

and so from (27.3), it follows (27.1) is equivalent to (27.2) for all $p$, including $p = s$. This proves the lemma.

**Corollary 27.1.7** The row reduced echelon form is unique in the sense that the algorithm described in Theorem 27.1.4 always leads to the same matrix.
Proof: Suppose $B$ and $C$ are both row reduced echelon forms for the matrix, $A$. Then they have the same zero columns since the above procedure does nothing to the zero columns. The pivot columns of $A$ are well defined by the above procedure and so $B$ and $C$ both have the sequence $e_1, e_2, \ldots$ occurring for the first time in the same positions, say in position, $i_1, i_2, \ldots, i_r$. Then from Lemma 27.1.6 every column in $B$ and $C$ between the $i_k$ and $i_{k+1}$ position is a linear combination involving the same scalars of the columns in the $i_1, \ldots, i_k$ position. This is equivalent to the assertion that each of these columns is identical and this proves the corollary.

Example 27.1.8 Find the row reduced echelon form of the matrix,

\[
\begin{pmatrix}
0 & 0 & 2 & 3 \\
0 & 2 & 0 & 1 \\
0 & 1 & 1 & 5
\end{pmatrix}
\]

The first nonzero column is the second in the matrix. We switch the third and first rows to obtain

\[
\begin{pmatrix}
0 & 1 & 1 & 5 \\
0 & 2 & 0 & 1 \\
0 & 0 & 2 & 3
\end{pmatrix}
\]

Now we multiply the top row by $-2$ and add to the second.

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

Next, add the second row to the bottom and then divide the bottom row by $-6$

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

Next use the bottom row to obtain zeros in the last column above the 1 and divide the second row by $-2$

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

Finally, add $-1$ times the middle row to the top.

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

This is in row reduced echelon form.

Example 27.1.9 Find the row reduced echelon form for the matrix,

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

You should verify that the row reduced echelon form is

\[
\begin{pmatrix}
1 & 0 & -\frac{2}{5} & 0 \\
0 & 1 & \frac{4}{5} & 1 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]
27.2 The Rank Of A Matrix

27.2.1 The Definition Of Rank

To begin, here is a definition to introduce some terminology.

**Definition 27.2.1** Let $A$ be an $m \times n$ matrix. The **column space** of $A$ is the subspace of $\mathbb{F}^m$ spanned by the columns. The **row space** is the subspace of $\mathbb{F}^n$ spanned by the rows.

There are three definitions of the rank of a matrix which are useful. These are given in the following definition. It turns out that the concept of determinant rank is the one most useful in applications to analysis but is virtually impossible to find directly. The other two concepts of rank are very easily determined and it is a happy fact that all three yield the same number.

**Definition 27.2.2** A **submatrix** of a matrix $A$ is a 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 space** of a matrix is the span of the rows and the **column space** of a matrix is the span of the columns. The row rank is the dimension of the row space and the column rank is the dimension of the column space. The rank of a matrix, $A$ is denoted by $\text{rank} (A)$.

**Example 27.2.3** Consider the matrix,

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

What is its rank?

You could look at all the $2 \times 2$ submatrices

\[
\begin{pmatrix}
1 & 2 \\
2 & 4
\end{pmatrix}, \begin{pmatrix}
1 & 3 \\
2 & 6
\end{pmatrix}, \begin{pmatrix}
2 & 3 \\
4 & 6
\end{pmatrix}
\]

Each has determinant equal to 0. Therefore, the rank is less than 2. Now look at the $1 \times 1$ submatrices. There exists one of these which has nonzero determinant. For example (1) has determinant equal to 1 and so the rank of this matrix equals 1.

Of course this example was pretty easy but what if you had a $4 \times 7$ matrix? You would have to consider all the $4 \times 4$ submatrices and then all the $3 \times 3$ submatrices and then all the $2 \times 2$ matrices and finally all the $1 \times 1$ matrices in order to compute the rank. Clearly this is not practical. The following theorem will remove the difficulties just indicated.

The following theorem is proved in the section on the theory of the determinant.

**Theorem 27.2.4** Let $A$ be an $m \times n$ matrix. Then the row rank, column rank and determinant rank are all the same.

27.2.2 Finding The Rank Of A Matrix

As indicated above, the row column, and determinant rank of a matrix are all the same. How do you compute the rank of a matrix? It turns out that this is an easy process if you use row operations.

**Corollary 27.2.5** Let $A$ and $B$ be two $m \times n$ matrices such that $B$ is obtained by applying a row operation to $A$. Then the two matrices have the same rank.
Proof: Suppose the column rank of $B$ is $r$. This means there are $r$ columns whose span yields all the columns of $B$. By Lemma 27.1.6 every column of $A$ is a linear combination of the corresponding columns in $A$. Therefore, the rank of $A$ is no larger than the rank of $B$. But $A$ may also be obtained from $B$ by a row operation. (Why?) Therefore, the same reasoning implies the rank of $B$ is no larger than the rank of $A$. This proves the corollary.

This suggests that to find the rank of a matrix, one should do row operations until a matrix is obtained in which its rank is obvious. The most obvious form from which to obtain the rank of a given matrix is the row reduced echelon form.

Example 27.2.6 Find the rank of the following matrix and find a basis for the column space and a basis for the row space.

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 1 & 3 & 6 & 0 & 2 \\ 3 & 7 & 8 & 6 & 6 \end{pmatrix}$$ (27.4)

Take $(-1)$ times the first row and add to the second and then take $(-3)$ times the first row and add to the third. This yields

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 0 & 1 & 5 & -3 & 0 \\ 0 & 1 & 5 & -3 & 0 \end{pmatrix}$$

By the above corollary, this matrix has the same rank as the first matrix. Now take $(-1)$ times the second row and add to the third row yielding

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 0 & 1 & 5 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Next take $(-2)$ times the second row and add to the first row. to obtain

$$\begin{pmatrix} 1 & 0 & -9 & 9 & 2 \\ 0 & 1 & 5 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$ (27.5)

Each of these row operations did not change the rank of the matrix. It is clear that linear combinations of the first two columns yield every other column so the rank of the matrix is no larger than 2. However, it is also clear that the determinant rank is at least 2 because, deleting every column other than the first two and every zero row yields the $2 \times 2$ identity matrix having determinant 1.

By Lemma 27.1.6 the first two columns of the original matrix yield all other columns as linear combinations. Also these first two columns must be linearly independent again by Lemma 27.1.6. Therefore, the first two columns of the original matrix are a basis for the column space of the original matrix. The two nonzero rows in the row reduced echelon form yield a basis for the row space of the original matrix. The reason is that the row operations do not take us out of the span of the rows of the original matrix so these two rows are in the row space and since all the row operations are reversible, it follows every row in the original row space is obtained as a linear combination of these two. The nonzero rows in the row reduced echelon form are linearly independent by construction.

Procedure 27.2.7 To find the rank of a matrix, obtain the row reduced echelon form for the matrix. Then count the number of nonzero rows. This is the rank. A basis for the row space is these nonzero rows and a basis for the column space of the original matrix are the pivot columns.
Example 27.2.8 Find the rank of the following matrix and give a basis for the row and column space.

\[
\begin{pmatrix}
1 & 2 & 1 & 3 & 2 \\
1 & 2 & 6 & 0 & 2 \\
3 & 6 & 8 & 6 & 6
\end{pmatrix}
\]

(27.6)

Take \((-1)\) times the first row and add to the second and then take \((-3)\) times the first row and add to the last row. This yields

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

Now multiply the second row by \(1/5\) and add 5 times it to the last row.

\[
\begin{pmatrix}
1 & 2 & 1 & 3 & 2 \\
0 & 0 & 1 & -3/5 & 0 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

(27.7)

Add \((-1)\) times the second row to the first.

\[
\begin{pmatrix}
1 & 2 & 0 & \frac{18}{5} & 2 \\
0 & 0 & 1 & -3/5 & 0 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

This is now in row reduced echelon form and so the rank is 2. A basis for the column space of the original matrix is

\[
\begin{pmatrix}
1 \\
1 \\
3
\end{pmatrix}, \begin{pmatrix}
1 \\
6 \\
8
\end{pmatrix}
\]

and a basis for the row space is \(\begin{pmatrix}
0 & 0 & 1 & -3/5 & 0
\end{pmatrix}, \begin{pmatrix}
1 & 2 & 0 & \frac{18}{5} & 2
\end{pmatrix}\).

The following corollary is a formal statement of the above procedure.

Corollary 27.2.9 The rank of a matrix equals the number of nonzero pivot columns. A basis for the column space is the pivot columns and a basis for the row space is the nonzero rows.

Proof: Row rank, determinant rank, and column rank are all the same so it suffices to consider only column rank. Write the row reduced echelon form for the matrix. From Corollary 27.2.5 this row reduced matrix has the same rank as the original matrix. Deleting all the zero rows and all the columns in the row reduced echelon form which do not correspond to a pivot column, yields an \(r \times r\) identity submatrix in which \(r\) is the number of pivot columns. Thus the rank is at least \(r\). Now from the construction of the row reduced echelon form, every column is a linear combination of these \(r\) columns. Therefore, the rank is no more than \(r\). The pivot columns in the row reduced echelon form are just \(e_1, \cdots, e_r\) where \(r\) is the rank, \(e_i\) being the \(i\) th column. Thus from the definition of row reduced echelon form, the columns, \(i_1, \cdots, i_r\) span all other columns in the row reduced echelon form and are linearly independent. Therefore, they are a basis for the column space of the row reduced echelon form of the given matrix. By Lemma 27.1.6 it follows the \(i_1, \cdots, i_r\) columns also are a basis for the column space of the original matrix.

The span of the nonzero rows gives the row space of the original matrix because they are obtained from row operations. These nonzero rows are also linearly independent and so they are a basis for the row space.
27.2.3 Rank And Existence Of Solutions To Linear Systems

Consider the linear system of equations,

$$Ax = b$$  \hspace{1cm} (27.8)

where $A$ is an $m \times n$ matrix, $x$ is a $n \times 1$ column vector, and $b$ is an $m \times 1$ column vector. Suppose

$$A = \begin{pmatrix} | & | \\
| & a_1 & \cdots & a_n |
| & | \\
\end{pmatrix}$$

where the $a_k$ denote the columns of $A$. Then if $x = (x_1, \cdots, x_n)^T$ is a solution of the system (27.8), it follows

$$x_1a_1 + \cdots + x_na_n = b$$

which says that $b$ is a vector in span $(a_1, \cdots, a_n)$. This shows that there exists a solution to the system, (27.8) if and only if $b$ is contained in span $(a_1, \cdots, a_n)$. In words, there is a solution to (27.8) if and only if $b$ is in the column space of $A$. In terms of rank, the following proposition describes the situation.

**Proposition 27.2.10** Let $A$ be an $m \times n$ matrix and let $b$ be an $m \times 1$ column vector. Then there exists a solution to (27.8) if and only if

$$\text{rank} \left( \begin{array}{c|c} A & b \end{array} \right) = \text{rank} (A).$$  \hspace{1cm} (27.9)

**Proof:** Place $\begin{pmatrix} A & b \end{pmatrix}$ and $A$ in row reduced echelon form, respectively $B$ and $C$. If the above condition on rank is true, then both $B$ and $C$ have the same number of nonzero rows. In particular, you cannot have a row of the form

$$\begin{pmatrix} 0 & \cdots & 0 & \Box \end{pmatrix}$$

where $\Box \neq 0$ in $B$. Therefore, there will exist a solution to the system.

Conversely, suppose there exists a solution. This means there cannot be such a row in $B$ described above. Therefore, $B$ and $C$ must have the same number of zero rows and so they have the same number of nonzero rows. Therefore, the rank of the two matrices in (27.9) is the same.

27.3 Exercises

1. Find the rank of the following matrices. If the rank is $r$, identify $r$ columns in the original matrix which have the property that every other column may be written as a linear combination of these. Also find a basis for the row and column spaces of the matrices.

(a) \[
\begin{pmatrix}
1 & 2 & 0 \\
3 & 2 & 1 \\
2 & 1 & 0 \\
0 & 2 & 1 
\end{pmatrix}
\]

(b) \[
\begin{pmatrix}
1 & 0 & 0 \\
4 & 1 & 1 \\
2 & 1 & 0 \\
0 & 2 & 0 
\end{pmatrix}
\]
27.3. EXERCISES

2. Determine which matrices are in row reduced echelon form.

(a) \[
\begin{pmatrix}
1 & 2 & 0 \\
0 & 1 & 7 \\
\end{pmatrix}
\]
(b) \[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & 0 & 0 \\
\end{pmatrix}
\]
(c) \[
\begin{pmatrix}
1 & 1 & 0 & 0 & 5 \\
0 & 0 & 1 & 2 & 0 \\
0 & 0 & 0 & 0 & 3 \\
\end{pmatrix}
\]

3. Row reduce the following matrices to obtain the row reduced echelon form. List the pivot columns in the original matrix.

(a) \[
\begin{pmatrix}
1 & 2 & 0 & 3 \\
2 & 1 & 2 & 2 \\
1 & 1 & 0 & 3 \\
\end{pmatrix}
\]
(b) \[
\begin{pmatrix}
1 & 2 & 3 \\
2 & 1 & -2 \\
3 & 0 & 0 \\
3 & 2 & 1 \\
\end{pmatrix}
\]
(c) \[
\begin{pmatrix}
1 & 2 & 1 & 3 \\
-3 & 2 & 1 & 0 \\
3 & 2 & 1 & 1 \\
\end{pmatrix}
\]

4. If \( \mathbf{b} \neq \mathbf{0} \), can the solution set of \( \mathbf{A} \mathbf{x} = \mathbf{b} \) be a plane through the origin? Explain.

5. Suppose \( \mathbf{A} \) is an \( m \times n \) matrix. Explain why the rank of \( \mathbf{A} \) is always no larger than \( \min(m,n) \).

6. Suppose \( \mathbf{A} \) is an \( m \times n \) matrix in which \( m \leq n \). Suppose also that the rank of \( \mathbf{A} \) equals \( m \). Show that \( \mathbf{A} \) maps \( \mathbb{F}^n \) onto \( \mathbb{F}^m \). Hint: The vectors \( \mathbf{e}_1, \ldots, \mathbf{e}_m \) occur as columns in the row reduced echelon form for \( \mathbf{A} \).

7. Suppose \( \mathbf{A} \) is an \( m \times n \) matrix in which \( m \geq n \). Suppose also that the rank of \( \mathbf{A} \) equals \( n \). Show that \( \mathbf{A} \) is one to one. Hint: If not, there exists a vector, \( \mathbf{x} \) such that \( \mathbf{A} \mathbf{x} = \mathbf{0} \), and this implies at least one column of \( \mathbf{A} \) is a linear combination of the others. Show this would require the column rank to be less than \( n \).
8. Explain why an \( n \times n \) matrix, \( A \) is both one to one and onto if and only if its rank is \( n \).

9. Suppose \( A \) is an \( m \times n \) matrix and \( \{w_1, \cdots, w_k\} \) is a linearly independent set of vectors in \( A(\mathbb{F}^n) \subseteq \mathbb{F}^m \). Now suppose \( A(z_i) = w_i \). Show \( \{z_1, \cdots, z_k\} \) is also independent.

10. Suppose \( A \) is an \( m \times n \) matrix and \( B \) is an \( n \times p \) matrix. Show that

\[
\dim (\ker (AB)) \leq \dim (\ker (A)) + \dim (\ker (B)).
\]

**Hint:** Consider the subspace, \( B(\mathbb{F}^p) \cap \ker (A) \) and suppose a basis for this subspace is \( \{w_1, \cdots, w_k\} \). Now suppose \( \{u_1, \cdots, u_r\} \) is a basis for \( \ker (B) \). Let \( \{z_1, \cdots, z_k\} \) be such that \( Bz_i = w_i \) and argue that \( \ker (AB) \subseteq \text{span} \{u_1, \cdots, u_r, z_1, \cdots, z_k\} \).

Here is how you do this. Suppose \( ABx = 0 \). Then \( Bx \in \ker (A) \cap B(\mathbb{F}^p) \) and so \( Bx = \sum_{i=1}^k Bz_i \) showing that

\[
x - \sum_{i=1}^k z_i \in \ker (B).
\]

11. Let \( H \) denote span \( \left( \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix} \right) \). Find the dimension of \( H \) and determine a basis.

12. Let \( H \) denote span \( \left( \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right) \). Find the dimension of \( H \) and determine a basis.

13. Let \( H \) denote span \( \left( \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 4 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right) \). Find the dimension of \( H \) and determine a basis.

14. Suppose a system of equations has fewer equations than variables and you have found a solution to this system of equations. Is it possible that your solution is the only one? Explain.

15. Suppose a system of linear equations has a \( 2 \times 4 \) augmented matrix and the last column is a pivot column. Could the system of linear equations be consistent? Explain.

16. Suppose the coefficient matrix of a system of \( n \) equations with \( n \) variables has the property that every column is a pivot column. Does it follow that the system of equations must have a solution? If so, must the solution be unique? Explain.

17. Suppose there is a unique solution to a system of linear equations. What must be true of the pivot columns in the augmented matrix.

18. State whether each of the following sets of data are possible for the matrix equation \( Ax = b \). If possible, describe the solution set. That is, tell whether there exists a unique solution no solution or infinitely many solutions.

(a) \( A \) is a \( 5 \times 6 \) matrix, \( \text{rank} (A) = 4 \) and \( \text{rank} (A|b) = 4 \). **Hint:** This says \( b \) is in the span of four of the columns. Thus the columns are not independent.
(b) $A$ is a $3 \times 4$ matrix, $\text{rank } (A) = 3$ and $\text{rank } (A|b) = 2$.

(c) $A$ is a $4 \times 2$ matrix, $\text{rank } (A) = 4$ and $\text{rank } (A|b) = 4$. **Hint:** This says $b$ is in the span of the columns and the columns must be independent.

(d) $A$ is a $5 \times 5$ matrix, $\text{rank } (A) = 4$ and $\text{rank } (A|b) = 5$. **Hint:** This says $b$ is not in the span of the columns.

(e) $A$ is a $4 \times 2$ matrix, $\text{rank } (A) = 2$ and $\text{rank } (A|b) = 2$. 
The \textit{LU} Decomposition

\subsection*{28.0.1 Outcomes}

1. Determine LU factorizations when possible.

2. Solve a linear system of equations using the LU decomposition.

\subsection*{28.1 Definition Of An \textit{LU} Decomposition}

An \textit{LU} decomposition of a matrix involves writing the given matrix as the product of a lower triangular matrix which has the main diagonal consisting entirely of ones \(L\), and an upper triangular matrix \(U\) in the indicated order. This is the version discussed here but it is sometimes the case that the \(L\) has numbers other than 1 down the main diagonal. It is still a useful concept. The \(L\) goes with “lower” and the \(U\) with “upper”. It turns out many matrices can be written in this way and when this is possible, people get excited about slick ways of solving the system of equations, \(Ax = y\). It is for this reason that you want to study the \textit{LU} decomposition. It allows you to work only with triangular matrices. It turns out that it takes about \(2n^3/3\) operations to use Gauss elimination but only \(n^3/3\) to obtain an \textit{LU} factorization.

First it should be noted not all matrices have an \textit{LU} decomposition and so we will emphasize the techniques for achieving it rather than formal proofs.

\textbf{Example 28.1.1} \textit{Can you write} \(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\) \textit{in the form LU as just described}?

To do so you would need
\[
\begin{pmatrix} 1 & 0 \\ x & 1 \end{pmatrix}\begin{pmatrix} a & b \\ 0 & c \end{pmatrix} = \begin{pmatrix} a & b \\ xa & xb+c \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.
\]

Therefore, \(b = 1\) and \(a = 0\). Also, from the bottom rows, \(xa = 1\) which can’t happen and have \(a = 0\). Therefore, you can’t write this matrix in the form \textit{LU}. It has no \textit{LU} decomposition. This is what we mean above by saying the method lacks generality.

\subsection*{28.2 Finding An \textit{LU} Decomposition}

Which matrices have an \textit{LU} decomposition? It turns out it is those whose row reduced echelon form can be achieved without switching rows and which only involve row operations of type 3 in which row \(j\) is replaced with a multiple of row \(i\) added to row \(j\) for \(i < j\).
Example 28.2.1 Find an LU decomposition of \( A = \begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 3 & 2 & 1 \\ 2 & 3 & 4 & 0 \end{pmatrix} \).

One way to find the LU decomposition is to simply look for it directly. You need
\[
\begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 3 & 2 & 1 \\ 2 & 3 & 4 & 0 \end{pmatrix} = \begin{pmatrix} a & d & h & j \\ x & 1 & 0 & 0 \\ y & z & 1 & 0 \end{pmatrix} \begin{pmatrix} a & d & h & j \\ 0 & b & e & i \\ 0 & 0 & c & f \end{pmatrix}.
\]

Then multiplying these you get
\[
\begin{pmatrix} a & d & h & j \\ xa & xd + b & xh + e & xj + i \\ ya & yd + zb & yh + ze + c & yj + iz + f \end{pmatrix}
\]
and so you can now tell what the various quantities equal. From the first column, you need \( a = 1, x = 1, y = 2 \). Now go to the second column. You need \( d = 2, xd + b = 3 \) so \( b = 1, yd + zb = 3 \) so \( z = -1 \). From the third column, \( h = 0, e = 2, c = 6 \). Now from the fourth column, \( j = 2, i = -1, f = -5 \). Therefore, an LU decomposition is
\[
\begin{pmatrix} 1 & 0 & 0 & 2 \\ 1 & 1 & 0 & 0 \\ 2 & -1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 & 2 \\ 0 & 1 & 2 & -1 \\ 0 & 0 & 6 & -5 \end{pmatrix}.
\]

You can check whether you got it right by simply multiplying these two.

There is also a convenient procedure for finding an LU decomposition. It turns out that it is only necessary to keep track of the multipliers which are used to row reduce to upper triangular form. This procedure is described in the following examples.

Example 28.2.2 Find an LU decomposition for \( A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & -4 \\ 1 & 5 & 2 \end{pmatrix} \).

Write the matrix next to the identity matrix as shown.
\[
\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & -4 \\ 1 & 5 & 2 \end{pmatrix}.
\]

The process involves doing row operations to the matrix on the right while simultaneously updating successive columns of the matrix on the left. First take \(-2\) times the first row and add to the second in the matrix on the right.
\[
\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -10 \\ 1 & 5 & 2 \end{pmatrix}
\]

Note the way we updated the matrix on the left. We put a 2 in the second entry of the first column because we used \(-2\) times the first row added to the second row. Now replace the third row in the matrix on the right by \(-1\) times the first row added to the third. Thus the next step is
\[
\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -10 \\ 0 & 3 & -1 \end{pmatrix}
\]
Finally, we will add the second row to the bottom row and make the following changes
\[
\begin{pmatrix}
1 & 0 & 0 \\
2 & 1 & 0 \\
1 & -1 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 2 & 3 \\
0 & -3 & -10 \\
0 & 0 & -11
\end{pmatrix}.
\]

At this point, we stop because the matrix on the right is upper triangular. An LU decomposition is the above.

The reason this gimmick works is related to the concept of products of elementary matrices of a certain special sort. It requires a little more development than we wish to consider here.

Example 28.2.3 Find an LU decomposition for \(A = \begin{pmatrix}
1 & 2 & 1 & 2 & 1 \\
2 & 0 & 2 & 1 & 1 \\
2 & 3 & 1 & 3 & 2 \\
1 & 0 & 1 & 1 & 2
\end{pmatrix}\).

We will use the same procedure as above. However, this time we will do everything for one column at a time. First multiply the first row by \((-1)\) and then add to the last row. Next take \((-2)\) times the first and add to the second and then \((-2)\) times the first and add to the third.

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

This finishes the first column of \(L\) and the first column of \(U\). Now take \(-1/4\) times the second row in the matrix on the right and add to the third followed by \(-1/2\) times the second added to the last.

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

This finishes the second column of \(L\) as well as the second column of \(U\). Since the matrix on the right is upper triangular, stop. The LU decomposition has now been obtained. This technique is called Dolittle’s method.

This process is entirely typical of the general case. The matrix \(U\) is just the first upper triangular matrix you come to in your quest for the row reduced echelon form using only the row operation which involves replacing a row by itself added to a multiple of another row. The matrix, \(L\) is what you get by updating the identity matrix as illustrated above.

You should note that for a square matrix, the number of row operations necessary to reduce to \(LU\) form is about half the number needed to place the matrix in row reduced echelon form. This is why an LU decomposition is of interest in solving systems of equations.

28.3 Solving Systems Using The LU Decomposition

The reason people care about the LU decomposition is it allows the quick solution of systems of equations. Here is an example.
Example 28.3.1 Suppose you want to find the solutions to \[ \begin{pmatrix} 1 & 2 & 3 & 2 \\ 4 & 3 & 1 & 1 \\ 1 & 2 & 3 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}. \]

Of course one way is to write the augmented matrix and grind away. However, this involves more row operations than the computation of the \( LU \) decomposition and it turns out that the \( LU \) decomposition can give the solution quickly. Here is how. The following is an \( LU \) decomposition for the matrix.

\[ \begin{pmatrix} 1 & 2 & 3 & 2 \\ 4 & 3 & 1 & 1 \\ 1 & 2 & 3 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -11 & -7 \\ 0 & 0 & 0 & -2 \end{pmatrix}. \]

Let \( Ux = y \) and consider \( Ly = b \) where in this case, \( b = (1, 2, 3)^T \). Thus

\[ \begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \]

which yields very quickly that \( y = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix} \). Now you can find \( x \) by solving \( Ux = y \). Thus in this case,

\[ \begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -11 & -7 \\ 0 & 0 & 0 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix} \]

which yields

\[ x = \begin{pmatrix} -\frac{3}{5} + \frac{7}{5}t \\ \frac{2}{5} - \frac{11}{5}t \\ t \\ -1 \end{pmatrix}, \quad t \in \mathbb{R}. \]

28.4 The \( PLU \) Decomposition

As indicated above, some matrices don’t have an \( LU \) decomposition. Here is an example.

\[ M = \begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 2 & 3 & 0 \\ 4 & 3 & 1 & 1 \end{pmatrix} \quad (28.1) \]

In this case, there is another decomposition which is useful called a \( PLU \) decomposition. Here \( P \) is a permutation matrix.

Example 28.4.1 Find a \( PLU \) decomposition for the above matrix in (28.1).
Proceed as before trying to find the row echelon form of the matrix. First add \(-1\) times the first row to the second row and then add \(-4\) times the first to the third. This yields
\[
\begin{pmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
4 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 2 & 3 & 2 \\
0 & 0 & 0 & -2 \\
0 & -5 & -11 & -7
\end{pmatrix}
\]
There is no way to do only row operations involving replacing a row with itself added to a multiple of another row to the matrix on the right in such a way as to obtain an upper triangular matrix. Therefore, consider the original matrix with the bottom two rows switched.

\[
M' = \begin{pmatrix}
1 & 2 & 3 & 2 \\
4 & 3 & 1 & 1 \\
1 & 2 & 3 & 0
\end{pmatrix}
= PM
\]

Now try again with this matrix. First take \(-1\) times the first row and add to the bottom row and then take \(-4\) times the first row and add to the second row. This yields
\[
\begin{pmatrix}
1 & 0 & 0 \\
4 & 1 & 0 \\
1 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 2 & 3 & 2 \\
0 & -5 & -11 & -7 \\
0 & 0 & 0 & -2
\end{pmatrix}
\]
The matrix on the right is upper triangular and so the \(LU\) decomposition of the matrix, \(M'\) has been obtained above.

Thus \(M' = PM = LU\) where \(L\) and \(U\) are given above. Notice that \(P^2 = I\) and therefore, \(M = P^2M = PLU\) and so
\[
\begin{pmatrix}
1 & 2 & 3 & 2 \\
1 & 2 & 3 & 0 \\
4 & 3 & 1 & 1
\end{pmatrix}
= \begin{pmatrix}
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & 1 & 0
\end{pmatrix}\begin{pmatrix}
1 & 0 & 0 \\
4 & 1 & 0 \\
1 & 0 & 1
\end{pmatrix}\begin{pmatrix}
1 & 2 & 3 & 2 \\
0 & -5 & -11 & -7 \\
0 & 0 & 0 & -2
\end{pmatrix}
\]

This process can always be followed and so there always exists a \(PLU\) decomposition of a given matrix even though there isn’t always an \(LU\) decomposition.

Example 28.4.2 Use the \(PLU\) decomposition of \(M\) where \(M = \begin{pmatrix}
1 & 2 & 3 \\
1 & 2 & 3 \\
4 & 3 & 1
\end{pmatrix}\) to solve the system \(Mx = b\) where \(b = (1, 2, 3)^T\).

Let \(Ux = y\) and consider \(PLy = b\). In other words, solve,
\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & 1 & 0
\end{pmatrix}
\begin{pmatrix}
1 & 0 & 0 \\
4 & 1 & 0 \\
1 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
y_1 \\
y_2 \\
y_3
\end{pmatrix}
= \begin{pmatrix}
1 \\
2 \\
3
\end{pmatrix}
\]

Multiplying both sides by \(P\) gives
\[
\begin{pmatrix}
1 & 0 & 0 \\
4 & 1 & 0 \\
1 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
y_1 \\
y_2 \\
y_3
\end{pmatrix}
= \begin{pmatrix}
1 \\
3 \\
2
\end{pmatrix}
\]
and so
\[
y = \begin{pmatrix}
y_1 \\
y_2 \\
y_3
\end{pmatrix}
= \begin{pmatrix}
1 \\
-1 \\
1
\end{pmatrix}
\]
THE LU DECOMPOSITION

Now $U \mathbf{x} = \mathbf{y}$ and so it only remains to solve

$$
\begin{pmatrix}
1 & 2 & 3 & 2 \\
0 & -5 & -11 & -7 \\
0 & 0 & 0 & -2
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{pmatrix}
= 
\begin{pmatrix}
1 \\
-1 \\
1
\end{pmatrix}
$$

which yields

$$
\begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{pmatrix}
= 
\begin{pmatrix}
\frac{1}{5} + \frac{7}{5} t \\
\frac{9}{5} - \frac{11}{5} t \\
t \\
-\frac{1}{2}
\end{pmatrix}
: t \in \mathbb{R}.
$$

28.5 Exercises

1. Find an LU decomposition of

$$
\begin{pmatrix}
1 & 2 & 0 \\
2 & 1 & 3 \\
1 & 2 & 3
\end{pmatrix}
$$

2. Find an LU decomposition of

$$
\begin{pmatrix}
1 & 2 & 3 & 2 \\
1 & 3 & 2 & 1 \\
5 & 0 & 1 & 3
\end{pmatrix}
$$

3. Find an LU decomposition of the matrix,

$$
\begin{pmatrix}
1 & -2 & -5 & 0 \\
-2 & 5 & 11 & 3 \\
3 & -6 & -15 & 1
\end{pmatrix}
$$

4. Find an LU decomposition of the matrix,

$$
\begin{pmatrix}
1 & -1 & -3 & -1 \\
-1 & 2 & 4 & 3 \\
2 & -3 & -7 & -3
\end{pmatrix}
$$

5. Find an LU decomposition of the matrix,

$$
\begin{pmatrix}
1 & -3 & -4 & -3 \\
-3 & 10 & 10 & 10 \\
1 & -6 & 2 & -5
\end{pmatrix}
$$

6. Find an LU decomposition of the matrix,

$$
\begin{pmatrix}
1 & 3 & 1 & -1 \\
3 & 10 & 8 & -1 \\
2 & 5 & -3 & -3
\end{pmatrix}
$$

7. Find an LU decomposition of the matrix,

$$
\begin{pmatrix}
3 & -2 & 1 \\
9 & -8 & 6 \\
-6 & 2 & -2 \\
3 & 2 & -7
\end{pmatrix}
$$

8. Find an LU decomposition of the matrix,

$$
\begin{pmatrix}
-3 & -1 & 3 \\
9 & 9 & -12 \\
3 & 19 & -16 \\
12 & 40 & -26
\end{pmatrix}
$$
9. Find an LU decomposition of the matrix, \( \begin{pmatrix} -1 & -3 & -1 \\ 1 & 3 & 0 \\ 3 & 9 & 0 \\ 4 & 12 & 16 \end{pmatrix} \).

10. Find the LU decomposition of the coefficient matrix using Dolittle’s method and use it to solve the system of equations.

   (a) \( x + 2y = 5 \)
   \( 2x + 3y = 6 \)
   \( x + 2y + z = 1 \)

   (b) \( 2x + 3y = 6 \)
   \( y + 3z = 2 \)
   \( x + 2y + 3z = 5 \)

   (c) \( 2x + 3y + z = 6 \)
   \( x - y + z = 2 \)
   \( x + 2y + 3z = 5 \)

   (d) \( 2x + 3y + z = 6 \)
   \( 3x + 5y + 4z = 11 \)

11. Find a PLU decomposition of \( \begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 2 \\ 2 & 1 & 1 \end{pmatrix} \).

12. Find a PLU decomposition of \( \begin{pmatrix} 1 & 2 & 1 & 2 & 1 \\ 2 & 4 & 2 & 4 & 1 \\ 1 & 2 & 1 & 3 & 2 \end{pmatrix} \).

13. Find a PLU decomposition of \( \begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 2 \\ 2 & 4 & 1 \\ 3 & 2 & 1 \end{pmatrix} \).

14. Is there only one LU decomposition for a given matrix? **Hint:** Consider the equation

\[
\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.
\]
Applications Of Spectral Theory

29.0.1 Outcomes

1. Find the principle directions of a deformation matrix.
2. Model a Markov process.
   (a) Find the limit state.
   (b) Determine comparisons of population after a long period of time.

29.1 Defective And Nondefective Matrices

Definition 29.1.1 By the fundamental theorem of algebra, it is possible to write the characteristic equation in the form

\[(\lambda - \lambda_1)^{r_1}(\lambda - \lambda_2)^{r_2} \cdots (\lambda - \lambda_m)^{r_m} = 0\]

where \(r_i\) is some integer no smaller than 1. Thus the eigenvalues are \(\lambda_1, \lambda_2, \ldots, \lambda_m\). The algebraic multiplicity of \(\lambda_j\) is defined to be \(r_j\).

Example 29.1.2 Consider the matrix,

\[A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \]  

(29.1)

What is the algebraic multiplicity of the eigenvalue \(\lambda = 1\)?

In this case the characteristic equation is

\[\det (A - \lambda I) = (1 - \lambda)^3 = 0\]

or equivalently,

\[\det (\lambda I - A) = (\lambda - 1)^3 = 0.\]

Therefore, \(\lambda\) is of algebraic multiplicity 3.

Definition 29.1.3 The geometric multiplicity of an eigenvalue is the dimension of the eigenspace,

\[\ker (A - \lambda I) = 0.\]

Example 29.1.4 Find the geometric multiplicity of \(\lambda = 1\) for the matrix in (29.1).
We need to solve
\[
\begin{pmatrix}
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= 
\begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}.
\]
The augmented matrix which must be row reduced to get this solution is therefore,
\[
\begin{pmatrix}
0 & 1 & 0 & | & 0 \\
0 & 0 & 1 & | & 0 \\
0 & 0 & 0 & | & 0
\end{pmatrix}
\]
This requires \(z = y = 0\) and \(x\) is arbitrary. Thus the eigenspace is
\[
t \begin{pmatrix}
1 \\
0 \\
0
\end{pmatrix}, \; t \in \mathbb{F}.
\]
It follows the geometric multiplicity of \(\lambda = 1\) is 1.

**Definition 29.1.5** An \(n \times n\) matrix is called **defective** if the geometric multiplicity is not equal to the algebraic multiplicity for some eigenvalue. Sometimes such an eigenvalue for which the geometric multiplicity is not equal to the algebraic multiplicity is called a defective eigenvalue. If the geometric multiplicity for an eigenvalue equals the algebraic multiplicity, the eigenvalue is sometimes referred to as nondefective.

Here is another more interesting example of a defective matrix.

**Example 29.1.6** Let
\[
A = \begin{pmatrix}
2 & -2 & -1 \\
-2 & -1 & -2 \\
14 & 25 & 14
\end{pmatrix}.
\]
Find the eigenvectors and eigenvalues.

In this case the eigenvalues are 3, 6, 6 where we have listed 6 twice because it is a zero of algebraic multiplicity two, the characteristic equation being
\[
(\lambda - 3)(\lambda - 6)^2 = 0.
\]
It remains to find the eigenvectors for these eigenvalues. First consider the eigenvectors for \(\lambda = 3\). You must solve
\[
\begin{pmatrix}
2 & -2 & -1 \\
-2 & -1 & -2 \\
14 & 25 & 14
\end{pmatrix}
- 3 \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= \begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}.
\]
The augmented matrix is
\[
\begin{pmatrix}
-1 & -2 & -1 & | & 0 \\
-2 & -4 & -2 & | & 0 \\
14 & 25 & 11 & | & 0
\end{pmatrix}
\]
and the row reduced echelon form is
\[
\begin{pmatrix}
1 & 0 & -1 & 0 \\
0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]
so the eigenvectors are nonzero vectors of the form
\[
\begin{pmatrix}
t \\
-t \\
t
\end{pmatrix} = t \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}
\]

Next consider the eigenvectors for \( \lambda = 6 \). This requires you to solve
\[
\begin{pmatrix}
  2 & -2 & -1 \\
-2 & -1 & -2 \\
 14 & 25 & 14
\end{pmatrix}
- 6
\begin{pmatrix}
  1 & 0 & 0 \\
 0 & 1 & 0 \\
 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
\]
and the augmented matrix for this system of equations is
\[
\begin{pmatrix}
  -4 & -2 & -1 & 0 \\
-2 & -7 & -2 & 0 \\
 14 & 25 & 8 & 0
\end{pmatrix}
\]
The row reduced echelon form is
\[
\begin{pmatrix}
  1 & 0 & \frac{1}{8} & 0 \\
 0 & 1 & \frac{1}{4} & 0 \\
 0 & 0 & 0 & 0
\end{pmatrix}
\]
and so the eigenvectors for \( \lambda = 6 \) are of the form
\[
t \begin{pmatrix}
  -\frac{1}{8} \\
  -\frac{1}{4} \\
  1
\end{pmatrix}
\]
or written more simply,
\[
t \begin{pmatrix}
  -1 \\
  -2 \\
  8
\end{pmatrix}
\]
where \( t \in \mathbb{F} \).

Note that in this example the eigenspace for the eigenvalue, \( \lambda = 6 \) is of dimension 1 because there is only one parameter. However, this eigenvalue is of multiplicity two as a root to the characteristic equation. Thus this eigenvalue is a defective eigenvalue. However, the eigenvalue 3 is nondefective. The matrix is defective because it has a defective eigenvalue.

The word, defective, seems to suggest there is something wrong with the matrix. This is in fact the case. Defective matrices are a lot of trouble in applications and we may wish they never occurred. However, they do occur as the above example shows. When you study linear systems of differential equations, you will have to deal with the case of defective matrices and you will see how awful they are. The reason these matrices are so horrible to work with is that it is impossible to obtain a basis of eigenvectors. When you study differential equations, solutions to first order systems are expressed in terms of eigenvectors of a certain matrix times \( e^{\lambda t} \) where \( \lambda \) is an eigenvalue. In order to obtain a general solution of this sort, you must have a basis of eigenvectors. For a defective matrix, such a basis does not exist and so you have to go to something called generalized eigenvectors. Unfortunately, it is never explained in beginning differential equations courses why there are enough generalized eigenvectors and eigenvectors to represent the general solution. In fact, this reduces to a difficult question in linear algebra equivalent to the existence of something
APPLICATIONS OF SPECTRAL THEORY

called the Jordan Canonical form which is much more difficult than everything discussed in the entire differential equations course. It is in the appendix called Heroic Linear Algebra.

Ultimately, the algebraic issues which will occur in differential equations are a red herring anyway. The real issues relative to existence of solutions to systems of ordinary differential equations are analytical, having much more to do with calculus than with linear algebra although this will likely not be made clear when you take a beginning differential equations class.

In terms of algebra, this lack of a basis of eigenvectors says that it is impossible to obtain a diagonal matrix which is similar to the given matrix.

Although there may be repeated roots to the characteristic equation, (10.2) and it is not known whether the matrix is defective in this case, there is an important theorem which holds when considering eigenvectors which correspond to distinct eigenvalues.

**Theorem 29.1.7** Suppose \( Mv_i = \lambda_i v_i, i = 1, \ldots, r, v_i \neq 0 \), and that if \( i \neq j \), then \( \lambda_i \neq \lambda_j \). Then the set of eigenvectors, \( \{v_1, \cdots, v_r\} \) is linearly independent.

**Proof:** If the conclusion of this theorem is not true, then there exist non zero scalars, \( c_{kj} \) such that

\[
\sum_{j=1}^{m} c_{kj} v_{kj} = 0. \tag{29.2}
\]

Take \( m \) to be the smallest number possible for an expression of the form (29.2) to hold. Then solving for \( v_{k1} \)

\[
v_{k1} = \sum_{k_j \neq k_1} d_{kj} v_{kj}, \tag{29.3}
\]

where \( d_{kj} = c_{kj}/c_{k1} \neq 0 \). Multiplying both sides by \( M \),

\[
\lambda_{k1} v_{k1} = \sum_{k_j \neq k_1} d_{kj} \lambda_{kj} v_{kj},
\]

which from (29.3) yields

\[
\sum_{k_j \neq k_1} d_{kj} \lambda_{k1} v_{kj} = \sum_{k_j \neq k_1} d_{kj} \lambda_{kj} v_{kj},
\]

and therefore,

\[
0 = \sum_{k_j \neq k_1} d_{kj} (\lambda_{k1} - \lambda_{kj}) v_{kj},
\]

a sum having fewer than \( m \) terms. However, from the assumption that \( m \) is as small as possible for (29.2) to hold with all the scalars, \( c_{k1} \) non zero, it follows that for some \( j \neq 1 \),

\[
d_{kj} (\lambda_{k1} - \lambda_{kj}) = 0
\]

which implies \( \lambda_{k1} = \lambda_{kj} \), a contradiction.

**29.2 Some Applications Of Eigenvalues And Eigenvectors**

**29.2.1 Principle Directions**

Recall that \( n \times n \) matrices can be considered as linear transformations. If \( F \) is a \( 3 \times 3 \) real matrix having positive determinant, it can be shown that \( F = RU \) where \( R \) is a rotation.
matrix and $U$ is a symmetric real matrix having positive eigenvalues. An application of this wonderful result, known to mathematicians as the **right polar decomposition**, is to continuum mechanics where a chunk of material is identified with a set of points in three dimensional space.

The linear transformation, $F$ in this context is called the **deformation gradient** and it describes the local deformation of the material. Thus it is possible to consider this deformation in terms of two processes, one which distorts the material and the other which just rotates it. It is the matrix, $U$ which is responsible for stretching and compressing. This is why in elasticity, the stress is often taken to depend on $U$ which is known in this context as the right **Cauchy Green strain tensor**. In this context, the eigenvalues will always be positive. The symmetry of $U$ allows the proof of a theorem which says that if $\lambda_M$ is the largest eigenvalue, then in every other direction other than the one corresponding to the eigenvector for $\lambda_M$ the material is stretched less than $\lambda_M$ and if $\lambda_m$ is the smallest eigenvalue, then in every other direction other than the one corresponding to an eigenvector of $\lambda_m$ the material is stretched more than $\lambda_m$. This process of writing a matrix as a product of two such matrices, one of which preserves distance and the other which distorts is also important in applications to geometric measure theory an interesting field of study in mathematics and to the study of quadratic forms which occur in many applications such as statistics. Here we are emphasizing the application to mechanics in which the eigenvectors of $U$ determine the **principle directions**, those directions in which the material is stretched the most or the least.

**Example 29.2.1** Find the principle directions determined by the matrix,

$$\begin{pmatrix}
\frac{29}{11} & \frac{6}{11} & \frac{6}{11} \\
\frac{6}{11} & \frac{41}{14} & \frac{19}{14} \\
\frac{6}{11} & \frac{19}{14} & \frac{41}{14}
\end{pmatrix}$$

The eigenvalues are $3$, $1$, and $\frac{1}{2}$.

It is nice to be given the eigenvalues. The largest eigenvalue is 3 which means that in the direction determined by the eigenvector associated with 3 the stretch is three times as large. The smallest eigenvalue is 1/2 and so in the direction determined by the eigenvector for 1/2 the material is stretched by a factor of 1/2, becoming locally half as long. It remains to find these directions. First consider the eigenvector for 3. It is necessary to solve

$$3 \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix} \begin{pmatrix}
x \\
y \\
z
\end{pmatrix} = \begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}$$

Thus the augmented matrix for this system of equations is

$$\begin{pmatrix}
\frac{4}{11} & -\frac{6}{11} & -\frac{6}{11} & | & 0 \\
-\frac{6}{11} & \frac{91}{14} & -\frac{49}{14} & | & 0 \\
-\frac{6}{11} & -\frac{19}{14} & \frac{91}{14} & | & 0
\end{pmatrix}$$
The row reduced echelon form is
\[
\begin{pmatrix}
1 & 0 & -3 & 0 \\
0 & 1 & -1 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]
and so the principle direction for the eigenvalue, 3 in which the material is stretched to the maximum extent is
\[
\begin{pmatrix}
3 \\
1 \\
1
\end{pmatrix}.
\]
A direction vector (or unit vector) in this direction is
\[
\begin{pmatrix}
3/\sqrt{11} \\
1/\sqrt{11} \\
1/\sqrt{11}
\end{pmatrix}.
\]
You should show that the direction in which the material is compressed the most is in the direction
\[
\begin{pmatrix}
0 \\
-1/\sqrt{2} \\
1/\sqrt{2}
\end{pmatrix}.
\]
Note this is meaningful information which you would have a hard time finding without the theory of eigenvectors and eigenvalues.

### 29.2.2 Migration Matrices

There are applications which are of great importance which feature only one eigenvalue.

**Definition 29.2.2** 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 \( \sum_i a_{ij} = 1 \). Such matrices in which the columns are nonnegative numbers which sum to one are called **Markov matrices**. In this context describing migration, they are also called **migration matrices**.

**Example 29.2.3** Here is an example of one of these matrices.
\[
\begin{pmatrix}
.4 & .2 \\
.6 & .8
\end{pmatrix}
\]
Thus if it is considered as a migration matrix, .4 is the proportion of residents in location 1 which stay in location one in a given time period while .6 is the proportion of residents in location 1 which move to location 2 and .2 is the proportion of residents in location 2 which move to location 1. Considered as a Markov matrix, these numbers are usually identified with probabilities.

If \( \mathbf{v} = (x_1, \ldots, x_n)^T \) 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 \( \sum_j a_{ij}x_j = (A\mathbf{v})_i \). Therefore, the population of location \( i \) after \( k \) years is \( (A^k\mathbf{v})_i \). An obvious application of this would be to a situation in which you rent trailers which can go to various parts of a city and you observe through experiments the proportion of trailers which go from point \( i \) to point \( j \) in a single day. Then you might want to find how many trailers would be in all the locations after 8 days.
Proposition 29.2.4 Let $A = (a_{ij})$ be a migration matrix. Then $1$ is always an eigenvalue for $A$.

**Proof:** Remember that $\det (B^T) = \det (B)$, Therefore, 
\[
\det (A - \lambda I) = \det (A - \lambda I)^T = \det (A^T - \lambda I)
\]
because $I^T = I$. Thus the characteristic equation for $A$ is the same as the characteristic equation for $A^T$ and so $A$ and $A^T$ have the same eigenvalues. We will show that $1$ is an eigenvalue for $A^T$ and then it will follow that $1$ is an eigenvalue for $A$.

Remember that for a migration matrix, $\sum_i a_{ij} = 1$. Therefore, if $A^T = (b_{ij})$ so $b_{ij} = a_{ji}$, it follows that
\[
\sum_j b_{ij} = \sum_j a_{ji} = 1.
\]

Therefore, from matrix multiplication,
\[
A^T \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \sum_j b_{1j} \\ \vdots \\ \sum_j b_{3j} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}
\]
which shows that \( \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \) is an eigenvector for $A^T$ corresponding to the eigenvalue, $\lambda = 1$.

As explained above, this shows that $\lambda = 1$ is an eigenvalue for $A$ because $A$ and $A^T$ have the same eigenvalues.

Example 29.2.5 Consider the migration matrix,
\[
\begin{pmatrix}
.6 & 0 & .1 \\
.2 & .8 & 0 \\
.2 & 2 & .9
\end{pmatrix}
\]
for locations 1, 2, and 3.

Suppose initially there are 100 residents in location 1, 200 in location 2 and 400 in location 4. Find the population in the three locations after 10 units of time.

From the above, it suffices to consider
\[
\begin{pmatrix}
.6 & 0 & .1 \\
.2 & .8 & 0 \\
.2 & 2 & .9
\end{pmatrix}^{10} \begin{pmatrix} 100 \\ 200 \\ 400 \end{pmatrix} = \begin{pmatrix}
115.08582922 \\
120.13067244 \\
464.78349834
\end{pmatrix}
\]
Of course you would need to round these numbers off.

A related problem asks for how many there will be in the various locations after a long time. It turns out that if some power of the migration matrix has all positive entries, then there is a limiting vector, $x = \lim_{k \to \infty} A^k x_0$ where $x_0$ is the initial vector describing the number of inhabitants in the various locations initially. This vector will be an eigenvector for the eigenvalue 1 because
\[
x = \lim_{k \to \infty} A^k x_0 = \lim_{k \to \infty} A^{k+1} x_0 = \lim_{k \to \infty} A^k x = Ax,
\]
and the sum of its entries will equal the sum of the entries of the initial vector, $x_0$ because this sum is preserved for every multiplication by $A$ since
\[
\sum_i \sum_j a_{ij} x_j = \sum_j x_j \left( \sum_i a_{ij} \right) = \sum_j x_j.
\]
Here is an example. It is the same example as the one above but here it will involve the long time limit.
Example 29.2.6 Consider the migration matrix,
\[
\begin{pmatrix}
0.6 & 0.1 \\
0.2 & 0.8 \\
0.2 & 0.9
\end{pmatrix}
\]
for locations 1, 2, and 3. Suppose initially there are 100 residents in location 1, 200 in location 2 and 400 in location 4. Find the population in the three locations after a long time.

You just need to find the eigenvector which goes with the eigenvalue 1 and then normalize it so the sum of its entries equals the sum of the entries of the initial vector. Thus you need to find a solution to
\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
- \begin{pmatrix}
0.6 & 0.1 \\
0.2 & 0.8 \\
0.2 & 0.9
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= \begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}
\]
The augmented matrix is
\[
\begin{pmatrix}
0.4 & 0 & -0.1 & 0 \\
-0.2 & 0.2 & 0 & 0 \\
-0.2 & -0.2 & 0.1 & 0
\end{pmatrix}
\]
and its row reduced echelon form is
\[
\begin{pmatrix}
1 & 0 & -0.25 & 0 \\
0 & 1 & -0.25 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]
Therefore, the eigenvectors are
\[
s \begin{pmatrix}
(1/4) \\
(1/4) \\
1
\end{pmatrix}
\]
and all that remains is to choose the value of s such that
\[
\frac{1}{4}s + \frac{1}{4}s + s = 100 + 200 + 400
\]
This yields $s = \frac{1400}{3}$ and so the long time limit would equal
\[
\frac{1400}{3} \begin{pmatrix}
(1/4) \\
(1/4) \\
1
\end{pmatrix}
= \begin{pmatrix}
116.66666666666667 \\
116.66666666666667 \\
466.66666666666667
\end{pmatrix}
\]
You would of course need to round these numbers off. You see that you are not far off after just 10 units of time. Therefore, you might consider this as a useful procedure because it is probably easier to solve a simple system of equations than it is to raise a matrix to a large power.

Example 29.2.7 Suppose a migration matrix is
\[
\begin{pmatrix}
\frac{1}{5} & \frac{1}{2} & \frac{1}{5} \\
\frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\
\frac{11}{20} & \frac{1}{4} & \frac{3}{20}
\end{pmatrix}
\]. Find the comparison between the populations in the three locations after a long time.
This amounts to nothing more than finding the eigenvector for \( \lambda = 1 \). Solve
\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
- \begin{pmatrix}
\frac{1}{5} & \frac{1}{2} & \frac{1}{5} \\
\frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\
\frac{11}{20} & \frac{1}{4} & \frac{3}{10}
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= \begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}.
\]
The augmented matrix is
\[
\begin{pmatrix}
\frac{4}{5} & -\frac{1}{2} & -\frac{1}{5} & | & 0 \\
-\frac{1}{4} & \frac{3}{4} & -\frac{1}{2} & | & 0 \\
-\frac{11}{20} & -\frac{1}{4} & \frac{7}{10} & | & 0
\end{pmatrix}
\]
The row echelon form is
\[
\begin{pmatrix}
1 & 0 & -\frac{16}{19} & 0 \\
0 & 1 & -\frac{18}{19} & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]
and so an eigenvector is
\[
\begin{pmatrix}
16 \\
18 \\
19
\end{pmatrix}.
\]
Thus there will be \( \frac{18}{19} \) more in location 2 than in location 1. There will be \( \frac{19}{18} \) more in location 3 than in location 2.

You see the eigenvalue problem makes these sorts of determinations fairly simple.

There are many other things which can be said about these sorts of migration problems. They include things like the gambler’s ruin problem which asks for the probability that a compulsive gambler will eventually lose all his money. However those problems are not so easy although they still involve eigenvalues and eigenvectors.

There are many other important applications of eigenvalue problems. We have just given a few such applications here. As pointed out, this is a very hard problem but sometimes you don’t need to find the eigenvalues exactly.

### 29.3 The Estimation Of Eigenvalues

There are ways to estimate the eigenvalues for matrices from just looking at the matrix. The most famous is known as Gershgorin’s theorem. This theorem gives a rough idea where the eigenvalues are just from looking at the matrix.

**Theorem 29.3.1** Let \( A \) be an \( n \times n \) matrix. Consider the \( n \) Gershgorin discs defined as
\[
D_i \equiv \left\{ \lambda \in \mathbb{C} : |\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\}.
\]
Then every eigenvalue is contained in some Gershgorin 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 \( \sigma(A) \), the spectrum of \( A \).

**Proof:** Suppose \( A\mathbf{x} = \lambda \mathbf{x} \) where \( \mathbf{x} \neq \mathbf{0} \). Then for \( A = (a_{ij}) \)

\[
\sum_{j \neq i} a_{ij} x_j = (\lambda - a_{ii}) x_i.
\]

Therefore, picking \( k \) such that \( |x_k| \geq |x_j| \) for all \( x_j \), it follows that \( |x_k| \neq 0 \) since \( |\mathbf{x}| \neq 0 \) and

\[
|x_k| \sum_{j \neq i} |a_{kj}| \geq \sum_{j \neq i} |a_{kj}| |x_j| \geq |\lambda - a_{ii}| |x_k|.
\]

Now dividing by \( |x_k| \), it follows \( \lambda \) is contained in the \( k^{th} \) Gerschgorin disc.

**Example 29.3.2** Suppose the matrix is

\[
A = \begin{pmatrix}
21 & -16 & -6 \\
14 & 60 & 12 \\
7 & 8 & 38
\end{pmatrix}
\]

Estimate the eigenvalues.

The exact eigenvalues are 35, 56, and 28. The Gerschgorin disks are

\[
D_1 = \{ \lambda \in \mathbb{C} : |\lambda - 21| \leq 22 \},
\]

\[
D_2 = \{ \lambda \in \mathbb{C} : |\lambda - 60| \leq 26 \},
\]

and

\[
D_3 = \{ \lambda \in \mathbb{C} : |\lambda - 38| \leq 15 \}.
\]

Gerschgorin’s theorem says these three disks contain the eigenvalues. Now 35 is in \( D_3 \), 56 is in \( D_2 \) and 28 is in \( D_1 \).

More can be said when the Gerschgorin disks are disjoint but this is an advanced topic which requires the theory of functions of a complex variable.

## 29.4 Exercises

1. State the eigenvalue problem from an algebraic perspective.

2. State the eigenvalue problem from a geometric perspective.

3. Suppose \( T \) is a linear transformation and it satisfies \( T^2 = T \) and \( T\mathbf{x} = \mathbf{x} \) for all \( \mathbf{x} \) in a certain subspace, \( V \). Show that 1 is an eigenvalue for \( T \) and show that all eigenvalues have absolute values no larger than 1.

4. Is it possible for a nonzero matrix to have only 0 as an eigenvalue?

5. Show that if \( A\mathbf{x} = \lambda \mathbf{x} \) and \( A\mathbf{y} = \lambda \mathbf{y} \), then whenever \( a, b \) are scalars,

\[
A (a\mathbf{x} + b\mathbf{y}) = \lambda (a\mathbf{x} + b\mathbf{y}).
\]

Does this imply that \( a\mathbf{x} + b\mathbf{y} \) is an eigenvector? Explain.
6. Let \( M \) be an \( n \times n \) matrix and suppose \( x_1, \ldots, x_n \) are \( n \) eigenvectors which form a linearly independent set. Form the matrix \( S \) by making the columns these vectors. Show that \( S^{-1} \) exists and that \( S^{-1}MS \) is a diagonal matrix (one having zeros everywhere except on the main diagonal) having the eigenvalues of \( M \) on the main diagonal. When this can be done the matrix is diagonalizable.

7. Show that a matrix, \( M \), is diagonalizable if and only if it has a basis of eigenvectors. **Hint:** The first part is done in Problem 6. It only remains to show that if the matrix can be diagonalized by some matrix, \( S \), giving \( D = S^{-1}MS \) for \( D \) a diagonal matrix, then it has a basis of eigenvectors. Try using the columns of the matrix \( S \).

8. Suppose \( A \) is an \( n \times n \) matrix which is diagonally dominant. This means

\[
|a_{ii}| > \sum_j |a_{ij}|
\]

Show that \( A^{-1} \) must exist.

9. Let \( M \) be an \( n \times n \) matrix. Then define the adjoint of \( M \), denoted by \( M^* \) to be the transpose of the conjugate of \( M \). For example,

\[
\begin{pmatrix}
2 & i \\
1 + i & 3
\end{pmatrix}^* = \begin{pmatrix}
2 & 1-i \\
-i & 3
\end{pmatrix}.
\]

A matrix, \( M \), is self adjoint if \( M^* = M \). Show the eigenvalues of a self adjoint matrix are all real. If the self adjoint matrix has all real entries, it is called symmetric. Show that the eigenvalues and eigenvectors of a symmetric matrix occur in conjugate pairs.

10. Suppose \( A \) is an \( n \times n \) matrix consisting entirely of real entries but \( a + ib \) is a complex eigenvalue having the eigenvector, \( x + iy \). Here \( x \) and \( y \) are real vectors. Show that then \( a - ib \) is also an eigenvalue with the eigenvector, \( x - iy \). **Hint:** You should remember that the conjugate of a product of complex numbers equals the product of the conjugates. Here \( a + ib \) is a complex number whose conjugate equals \( a - ib \).

11. Recall an \( n \times n \) matrix is said to be symmetric if it has all real entries and if \( A = A^T \). Show the eigenvectors and eigenvalues of a real symmetric matrix are real.

12. Recall an \( n \times n \) matrix is said to be skew symmetric if it has all real entries and if \( A = -A^T \). Show that any nonzero eigenvalues must be of the form \( ib \) where \( i^2 = -1 \). In words, the eigenvalues are either 0 or pure imaginary. Show also that the eigenvectors corresponding to the pure imaginary eigenvalues are imaginary in the sense that every entry is of the form \( ix \) for \( x \in \mathbb{R} \).

13. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
-19 & -14 & -1 \\
8 & 4 & 8 \\
15 & 30 & -3
\end{pmatrix}.
\]

Determine whether the matrix is defective.

14. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
-3 & -30 & 15 \\
0 & 12 & 0 \\
15 & 30 & -3
\end{pmatrix}.
\]

Determine whether the matrix is defective.
15. Find the eigenvalues and eigenvectors of the matrix
\[
\begin{pmatrix}
8 & 4 & 5 \\
0 & 12 & 9 \\
-2 & 2 & 10
\end{pmatrix}.
\]
Determine whether the matrix is defective.

16. Find the eigenvalues and eigenvectors of the matrix
\[
\begin{pmatrix}
7 & -2 & 0 \\
8 & -1 & 0 \\
-2 & 4 & 6
\end{pmatrix}.
\]
Can you find three independent eigenvectors?

17. Find the eigenvalues and eigenvectors of the matrix
\[
\begin{pmatrix}
3 & -2 & -1 \\
0 & 5 & 1 \\
0 & 2 & 4
\end{pmatrix}.
\]
Can you find three independent eigenvectors in this case?

18. Find the eigenvalues and eigenvectors of the matrix
\[
\begin{pmatrix}
12 & -12 & 6 \\
0 & 18 & 0 \\
6 & 12 & 12
\end{pmatrix}.
\]

19. Find the eigenvalues and eigenvectors of the matrix
\[
\begin{pmatrix}
-5 & -1 & 10 \\
-15 & 9 & -6 \\
8 & -8 & 2
\end{pmatrix}.
\]
Determine whether the matrix is defective.

20. Find the eigenvalues and eigenvectors of the matrix
\[
\begin{pmatrix}
-10 & -8 & 8 \\
-4 & -14 & -4 \\
0 & 0 & -18
\end{pmatrix}.
\]
Determine whether the matrix is defective.

21. Find the eigenvalues and eigenvectors of the matrix
\[
\begin{pmatrix}
1 & 26 & -17 \\
4 & -4 & 4 \\
-9 & -18 & 9
\end{pmatrix}.
\]
Determine whether the matrix is defective.

22. Find the eigenvalues and eigenvectors of the matrix
\[
\begin{pmatrix}
8 & 4 & 5 \\
0 & 12 & 9 \\
-2 & 2 & 10
\end{pmatrix}.
\]
Determine whether the matrix is defective.
23. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} 9 & 6 & -3 \\ 0 & 6 & 0 \\ -3 & -6 & 9 \end{pmatrix}.$$ Determine whether the matrix is defective.

24. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} -10 & -2 & 11 \\ -18 & 6 & -9 \\ 10 & -10 & -2 \end{pmatrix}.$$ Determine whether the matrix is defective.

25. Find the complex eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} 4 & -2 & -2 \\ 0 & 2 & -2 \\ 2 & 0 & 2 \end{pmatrix}.$$ Determine whether the matrix is defective.

26. Find the complex eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} -4 & 2 & 0 \\ 2 & -4 & 0 \\ -2 & 2 & -2 \end{pmatrix}.$$ Determine whether the matrix is defective.

27. Find the complex eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} 1 & 1 & -6 \\ 7 & -5 & -6 \\ -1 & 7 & 2 \end{pmatrix}.$$ Determine whether the matrix is defective.

28. Find the complex eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} 4 & 2 & 0 \\ -2 & 4 & 0 \\ -2 & 2 & 6 \end{pmatrix}.$$ Determine whether the matrix is defective.

29. Let $T$ be the linear transformation which reflects vectors about the $x$ axis. Find a matrix for $T$ and then find its eigenvalues and eigenvectors.

30. Let $T$ be the linear transformation which rotates all vectors in $\mathbb{R}^2$ counterclockwise through an angle of $\pi/2$. Find a matrix of $T$ and then find eigenvalues and eigenvectors.

31. Let $T$ be the linear transformation which reflects all vectors in $\mathbb{R}^3$ through the $xy$ plane. Find a matrix for $T$ and then obtain its eigenvalues and eigenvectors.

32. You own a trailer rental company in a large city and you have four locations, one in the South East, one in the North East, one in the North West, and one in the South West. Denote these locations by SE, NE, NW, and SW respectively. Suppose you observe that in a typical day, .8 of the trailers starting in SE stay in SE, .1 of the trailers in NE go to SE, .1 of the trailers in NW end up in SE, .2 of the trailers in SW end up in SE, .1 of the trailers in NE end up in NE, .7 of the trailers in NE end up in NE, .2 of the trailers in SW end up in NE, .1 of the trailers in SE end up in SW, .5 of the trailers in SW end up in SW, .2 of the trailers in SW end up in NW, .1 of the trailers in SW end up in SW, .6 of the trailers in SW end up in SW, .1 of the trailers in SW end up in SW, .5 of the trailers in SW end up in SW. You begin with 20 trailers in each location. Approximately how many will you have in each location after a long time? Will any location ever run out of trailers?
33. Suppose the migration matrix for three locations is
\[
\begin{pmatrix}
0.5 & 0 & 0.3 \\
0.3 & 0.8 & 0 \\
0.2 & 0.2 & 0.7
\end{pmatrix}.
\]
Find a comparison for the populations in the three locations after a long time.

34. Suppose the migration matrix for three locations is
\[
\begin{pmatrix}
0.1 & 0.1 & 0.3 \\
0.3 & 0.7 & 0 \\
0.6 & 0.2 & 0.7
\end{pmatrix}.
\]
Find a comparison for the populations in the three locations after a long time.

35. Find the principle direction for stretching for the matrix,
\[
\begin{pmatrix}
13 & 2\sqrt{5} & 8\sqrt{5} \\
\frac{2\sqrt{5}}{5} & 6 & 4 \\
\frac{8\sqrt{5}}{15} & \frac{4}{15} & \frac{64}{45}
\end{pmatrix}.
\]
The eigenvalues are 2 and 1.

36. Find the principle directions for the matrix,
\[
\begin{pmatrix}
\frac{5}{2} & -\frac{1}{2} & 0 \\
-\frac{1}{2} & \frac{5}{2} & 0 \\
0 & 0 & 1
\end{pmatrix}.
Some Special Matrices

30.0.1 Outcomes

1. Define symmetric matrix, skew-symmetric matrix, and orthogonal matrix. Prove identities involving these types of matrices.

2. Characterize and determine the eigenvalues and eigenvectors of symmetric, skew-symmetric, and orthogonal matrices. Derive basic facts concerning these matrices.

3. Define an orthonormal set of vectors. Determine whether a set of vectors is orthonormal.

4. Relate the orthogonality of a matrix to the orthonormality of its column (or row) vectors.

5. Diagonalize a symmetric matrix.

30.1 Symmetric And Orthogonal Matrices

30.1.1 Orthogonal Matrices

Remember that to find the inverse of a matrix was often a long process. However, it was very easy to take the transpose of a matrix. For some matrices, the transpose equals the inverse and when the matrix has all real entries, and this is true, it is called an orthogonal matrix.

Definition 30.1.1 A real \( n \times n \) matrix, \( U \) is called an orthogonal matrix if \( UU^T = U^T U = I \). Thus an orthogonal matrix is a unitary matrix which happens to be real.

Example 30.1.2 Show the matrix,

\[
U = \begin{pmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}
\end{pmatrix}
\]

is orthogonal.

\[
UU^T = \begin{pmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}
\end{pmatrix} \begin{pmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}
\end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.
\]
Example 30.1.3 Let \( U = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} \). Is \( U \) orthogonal?

The answer is yes. This is because the columns form an orthonormal set of vectors as well as the rows. As discussed above this is equivalent to \( U^T U = I \).

\[
U^T U = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}^T \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}
\]

When you say that \( U \) is orthogonal, you are saying that \( \sum_j U_{ij} U_{jk} = \sum_j U_{ij} U_{kj} = \delta_{ik} \).

In words, the dot product of the \( i \)th row of \( U \) with the \( k \)th row gives 1 if \( i = k \) and 0 if \( i \neq k \). The same is true of the columns because \( U^T U = I \) also. Therefore,

\[
\sum_j U^T_{ij} U_{jk} = \sum_j U_{ji} U_{jk} = \delta_{ik}
\]

which says that the one column dotted with another column gives 1 if the two columns are the same and 0 if the two columns are different.

More succinctly, this states that if \( u_1, \cdots, u_n \) are the columns of \( U \), an orthogonal matrix, then

\[
u_i \cdot u_j = \delta_{ij} \equiv \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}
\]

Definition 30.1.4 A set of vectors, \( \{u_1, \cdots, u_n\} \) is said to be an orthonormal set if (30.1).

Theorem 30.1.5 If \( \{u_1, \cdots, u_n\} \) is an orthonormal set of vectors then it is linearly independent.

Proof: Using the properties of the dot product,

\[
0 \cdot u = (0 + 0) \cdot u = 0 \cdot u + 0 \cdot u
\]

and so, subtracting \( 0 \cdot u \) from both sides yields \( 0 \cdot u = 0 \). Now suppose \( \sum_j c_j u_j = 0 \). Then from the properties of the dot product,

\[
c_k = \sum_j c_j \delta_{jk} = \sum_j c_j (u_j \cdot u_k) = \left( \sum_j c_j u_j \right) \cdot u_k = 0 \cdot u_k = 0.
\]

Since \( k \) was arbitrary, this shows that each \( c_k = 0 \) and this has shown that if \( \sum_j c_j u_j = 0 \), then each \( c_j = 0 \). This is what it means for the set of vectors to be linearly independent.

Example 30.1.6 Let \( U = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{\sqrt{3}}{3} \end{pmatrix} \). Is \( U \) an orthogonal matrix?
The answer is yes. This is because the columns (rows) form an orthonormal set of vectors.

The importance of orthogonal matrices is that they change components of vectors relative to different Cartesian coordinate systems. Geometrically, the orthogonal matrices are exactly those which preserve all distances in the sense that if \( x \in \mathbb{R}^n \) and \( U \) is orthogonal, then \( ||Ux|| = ||x|| \) because

\[
||Ux||^2 = (Ux)^T Ux = x^T U^T Ux = x^T Ix = ||x||^2.
\]

**Observation 30.1.7** Suppose \( U \) is an orthogonal matrix. Then \( \det (U) = \pm 1 \).

This is easy to see from the properties of determinants. Thus

\[
\det (U)^2 = \det (U^T) \det (U) = \det (U^T U) = \det (I) = 1.
\]

Orthogonal matrices are divided into two classes, proper and improper. The proper orthogonal matrices are those whose determinant equals 1 and the improper ones are those whose determinants equal -1. The reason for the distinction is that the improper orthogonal matrices are sometimes considered to have no physical significance since they cause a change in orientation which would correspond to material passing through itself in a non physical manner. Thus in considering which coordinate systems must be considered in certain applications, you only need to consider those which are related by a proper orthogonal transformation. Geometrically, the linear transformations determined by the proper orthogonal matrices correspond to the composition of rotations.

### 30.1.2 Symmetric And Skew Symmetric Matrices

**Definition 30.1.8** A real \( n \times n \) matrix, \( A \), is **symmetric** if \( A^T = A \). If \( A = -A^T \), then \( A \) is called **skew symmetric**.

**Theorem 30.1.9** The eigenvalues of a real symmetric matrix are real. The eigenvalues of a real skew symmetric matrix are 0 or pure imaginary.

**Proof:** The proof is best understood as a special case of more general considerations. However, here is a proof in this special case.

Recall that for a complex number, \( a + ib \), the complex conjugate, denoted by \( \overline{a + ib} \) is given by the formula \( \overline{a + ib} = a - ib \). The notation, \( \overline{x} \) will denote the vector which has every entry replaced by its complex conjugate.

Suppose \( A \) is a real symmetric matrix and \( Ax = \lambda x \). Then

\[
\overline{Ax}^T x = (Ax)^T x = \overline{x}^T A^T x = \overline{x}^T Ax = \lambda \overline{x}^T x.
\]

Dividing by \( \overline{x}^T x \) on both sides yields \( \overline{A} = \lambda \) which says \( \lambda \) is real. (Why?)

Next suppose \( A = -A^T \) so \( A \) is skew symmetric and \( Ax = \lambda x \). Then

\[
\overline{Ax}^T x = (Ax)^T x = \overline{x}^T A^T x = -\overline{x}^T Ax = -\lambda \overline{x}^T x
\]

and so, dividing by \( \overline{x}^T x \) as before, \( \overline{A} = -\lambda \). Letting \( \lambda = a + ib \), this means \( a - ib = -a - ib \) and so \( a = 0 \). Thus \( \lambda \) is pure imaginary.

**Example 30.1.10** Let \( A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \). This is a skew symmetric matrix. Find its eigenvalues.
Its eigenvalues are obtained by solving the equation \( \det \begin{pmatrix} -\lambda & -1 \\ 1 & -\lambda \end{pmatrix} = \lambda^2 + 1 = 0 \). You see the eigenvalues are \( \pm i \), pure imaginary.

**Example 30.1.11** Let \( A = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix} \). This is a symmetric matrix. Find its eigenvalues.

Its eigenvalues are obtained by solving the equation, \( \det \begin{pmatrix} 1 - \lambda & 2 \\ 2 & 3 - \lambda \end{pmatrix} = -1 - 4\lambda + \lambda^2 = 0 \) and the solution is \( \lambda = 2 + \sqrt{5} \) and \( \lambda = 2 - \sqrt{5} \).

**Definition 30.1.12** An \( n \times n \) matrix, \( A = (a_{ij}) \) is called a diagonal matrix if \( a_{ij} = 0 \) whenever \( i \neq j \). For example, a diagonal matrix is of the form indicated below where * denotes a number.

\[
\begin{pmatrix}
\ast & 0 & \cdots & 0 \\
0 & \ast & & \\
& \ddots & \ddots & 0 \\
0 & \cdots & 0 & \ast
\end{pmatrix}
\]

**Theorem 30.1.13** Let \( A \) be a real symmetric matrix. Then there exists an orthogonal matrix, \( U \) such that \( U^T A U \) is a diagonal matrix. Moreover, the diagonal entries are the eigenvalues of \( A \).

**Proof:** This is proved in Corollary 10.4.9 on Page 197.

**Corollary 30.1.14** If \( A \) is a real \( n \times n \) symmetric matrix, then there exists an orthonormal set of eigenvectors, \( \{u_1, \ldots, u_n\} \).

**Proof:** Since \( A \) is symmetric, then by Theorem 30.1.13, there exists an orthogonal matrix, \( U \) such that \( U^T A U \) is a diagonal matrix whose diagonal entries are the eigenvalues of \( A \). Therefore, since \( A \) is symmetric and all the matrices are real,

\[
D = D^T = U^T A^T U = U^T A^T U = U^T A U = D
\]

showing \( D \) is real because each entry of \( D \) equals its complex conjugate.\(^1\)

Finally, let

\[
U = \begin{pmatrix}
u_1 & u_2 & \cdots & u_n
\end{pmatrix}
\]

where the \( u_i \) denote the columns of \( U \) and

\[
D = \begin{pmatrix}
\lambda_1 & 0 \\
& \ddots & \ddots & \\
0 & & \lambda_n
\end{pmatrix}
\]

The equation, \( U^T A U = D \) 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 orthogonal, the \( ij^{th} \) entry of \( U^T U \) equals \( \delta_{ij} \) and so

\[
\delta_{ij} = u_i^T u_j = u_i \cdot u_j.
\]

This proves the corollary because it shows the vectors \( \{u_i\} \) form an orthonormal basis.

The following corollary is also important.

\(^1\)Recall that for a complex number, \( x + iy \), the complex conjugate, denoted by \( \overline{x + iy} \), is defined as \( x - iy \).
Example 30.1.15 Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix,

\[
\begin{pmatrix}
\frac{19}{9} & -\frac{8}{15}\sqrt{5} & -\frac{2}{45}\sqrt{5} \\
-\frac{8}{15}\sqrt{5} & -\frac{1}{5} & -\frac{16}{15} \\
-\frac{2}{45}\sqrt{5} & -\frac{16}{15} & \frac{94}{45}
\end{pmatrix}
\]

given that the eigenvalues are 3, -1, and 2.

The augmented matrix which needs to be row reduced to find the eigenvectors for \(\lambda = 3\) is

\[
\begin{pmatrix}
\frac{19}{9} - 3 & -\frac{8}{15}\sqrt{5} & -\frac{2}{45}\sqrt{5} & | & 0 \\
\frac{8}{15}\sqrt{5} & -\frac{1}{5} - 3 & -\frac{16}{15} & | & 0 \\
\frac{2}{45}\sqrt{5} & -\frac{16}{15} & -\frac{94}{45} & | & 0
\end{pmatrix}
\]

and the row reduced echelon form for this is

\[
\begin{pmatrix}
1 & 0 & -\frac{1}{2}\sqrt{5} & | & 0 \\
0 & 1 & \frac{3}{4} & | & 0 \\
0 & 0 & 0 & | & 0
\end{pmatrix}
\]

Therefore, eigenvectors for \(\lambda = 3\) are

\[
z \begin{pmatrix} \frac{1}{2}\sqrt{5} \\ -\frac{3}{4} \\ 1 \end{pmatrix}
\]

where \(z \neq 0\).

The augmented matrix which must be row reduced to find the eigenvectors for \(\lambda = -1\) is

\[
\begin{pmatrix}
\frac{19}{9} + 1 & -\frac{8}{15}\sqrt{5} & -\frac{2}{45}\sqrt{5} & | & 0 \\
\frac{8}{15}\sqrt{5} & -\frac{1}{5} + 1 & -\frac{16}{15} & | & 0 \\
\frac{2}{45}\sqrt{5} & -\frac{16}{15} & \frac{94}{45} + 1 & | & 0
\end{pmatrix}
\]

and the row reduced echelon form is

\[
\begin{pmatrix}
1 & 0 & -\frac{1}{2}\sqrt{5} & | & 0 \\
0 & 1 & -3 & | & 0 \\
0 & 0 & 0 & | & 0
\end{pmatrix}
\]

Therefore, the eigenvectors for \(\lambda = -1\) are

\[
z \begin{pmatrix} \frac{1}{2}\sqrt{5} \\ 3 \\ 1 \end{pmatrix}, \quad z \neq 0
\]
The augmented matrix which must be row reduced to find the eigenvectors for \( \lambda = 2 \) is
\[
\begin{pmatrix}
\frac{10}{7} - 2 & -\frac{8}{7} \sqrt{5} & \frac{2}{7} \sqrt{5} & 0 \\
-\frac{8}{7} \sqrt{5} & -\frac{1}{5} - 2 & -\frac{16}{7} & 0 \\
\frac{2}{7} \sqrt{5} & -\frac{16}{7} & \frac{94}{7} - 2 & 0
\end{pmatrix}
\]
and its row reduced echelon form is
\[
\begin{pmatrix}
1 & 0 & \frac{2}{7} \sqrt{5} & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]
so the eigenvectors for \( \lambda = 2 \) are
\[
z \begin{pmatrix}
-\frac{2}{7} \sqrt{5} \\
0 \\
1
\end{pmatrix}, \quad z \neq 0.
\]

It remains to find an orthonormal basis. You can check that the dot product of any of these vectors with another of them gives zero and so it suffices choose \( z \) in each case such that the resulting vector has length 1. First consider the vectors for \( \lambda = 3 \). It is required to choose \( z \) such that
\[
z \begin{pmatrix}
\frac{1}{3} \sqrt{5} \\
-\frac{3}{4}
\end{pmatrix}
\]
is a unit vector. In other words, you need
\[
z \begin{pmatrix}
\frac{1}{3} \sqrt{5} \\
-\frac{3}{4}
\end{pmatrix} \cdot z \begin{pmatrix}
\frac{1}{3} \sqrt{5} \\
-\frac{3}{4}
\end{pmatrix} = 1.
\]
But the above dot product equals \( \frac{45}{16} z^2 \) and this equals 1 when \( z = \frac{4}{15} \sqrt{5} \). Therefore, the eigenvector which is desired is
\[
\frac{4}{15} \sqrt{5} \begin{pmatrix}
\frac{1}{3} \sqrt{5} \\
-\frac{3}{4}
\end{pmatrix} = \begin{pmatrix}
\frac{2}{7} \sqrt{5} \\
-\frac{1}{5} \sqrt{5}
\end{pmatrix}.
\]

Next find the eigenvector for \( \lambda = -1 \). The same process requires that \( 1 = \frac{45}{16} z^2 \) which happens when \( z = \frac{2}{15} \sqrt{5} \). Therefore, an eigenvector for \( \lambda = -1 \) which has unit length is
\[
\frac{2}{15} \sqrt{5} \begin{pmatrix}
\frac{1}{2} \sqrt{5} \\
3
\end{pmatrix} = \begin{pmatrix}
\frac{1}{2} \sqrt{5} \\
\frac{2}{7} \sqrt{5}
\end{pmatrix}.
\]
Finally, consider $\lambda = 2$. This time you need $1 = \frac{9}{5}z^2$ which occurs when $z = \frac{1}{3}\sqrt{5}$.

Therefore, the eigenvector is

$$\frac{1}{3}\sqrt{5} \begin{pmatrix} -\frac{2}{5}\sqrt{5} \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -\frac{2}{3} \\ 0 \\ \frac{1}{3}\sqrt{5} \end{pmatrix}.$$  

Now recall that the vectors form an orthonormal set of vectors if the matrix having them as columns is orthogonal. That matrix is

$$\begin{pmatrix} \frac{2}{5} & \frac{1}{3} & -\frac{2}{3} \\ -\frac{1}{5}\sqrt{5} & \frac{2}{5}\sqrt{5} & 0 \\ \frac{1}{15}\sqrt{5} & \frac{2}{15}\sqrt{5} & \frac{1}{3}\sqrt{5} \end{pmatrix}.$$  

Is this orthogonal? To find out, multiply by its transpose. Thus

$$\begin{pmatrix} \frac{2}{5} & -\frac{1}{5}\sqrt{5} & \frac{1}{3}\sqrt{5} \\ \frac{1}{3} & \frac{2}{5}\sqrt{5} & \frac{2}{15}\sqrt{5} \\ -\frac{2}{5} & 0 & \frac{1}{3}\sqrt{5} \end{pmatrix} \begin{pmatrix} \frac{2}{3} \\ -\frac{1}{5}\sqrt{5} \\ \frac{1}{15}\sqrt{5} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$  

Since the identity was obtained this shows the above matrix is orthogonal and that therefore, the columns form an orthonormal set of vectors. The problem asks for you to find an orthonormal set of vectors in $\mathbb{R}^n$ is always a basis. Therefore, since there are three of these vectors, they must constitute a basis.

**Example 30.1.16** Find an orthonormal set of three eigenvectors for the matrix,

$$\begin{pmatrix} \frac{13}{9} & \frac{2}{15}\sqrt{5} & \frac{6}{15}\sqrt{5} \\ \frac{2}{15}\sqrt{5} & \frac{6}{5} & \frac{4}{15} \\ \frac{8}{15}\sqrt{5} & \frac{4}{15} & \frac{61}{15} \end{pmatrix}$$  

given the eigenvalues are 2, and 1.

The eigenvectors which go with $\lambda = 2$ are obtained from row reducing the matrix

$$\begin{pmatrix} \frac{13}{9} & \frac{2}{15}\sqrt{5} & \frac{6}{15}\sqrt{5} & | & 0 \\ \frac{2}{15}\sqrt{5} & \frac{6}{5} & \frac{4}{15} & | & 0 \\ \frac{8}{15}\sqrt{5} & \frac{4}{15} & \frac{61}{15} & -2 & | & 0 \end{pmatrix}$$  

and its row reduced echelon form is

$$\begin{pmatrix} 1 & -\frac{1}{2}\sqrt{5} & | & 0 \\ 0 & 1 & -\frac{2}{3} & | & 0 \\ 0 & 0 & 0 & | & 0 \end{pmatrix}.$$
which shows the eigenvectors for $\lambda = 2$ are

$$z \begin{pmatrix} \frac{1}{2} \sqrt{5} \\ \frac{3}{4} \\ 1 \end{pmatrix}$$

and a choice for $z$ which will produce a unit vector is $z = \frac{4}{15} \sqrt{5}$. Therefore, the vector we want is

$$\begin{pmatrix} \frac{2}{3} \\ \frac{1}{5} \sqrt{5} \\ \frac{4}{15} \sqrt{5} \end{pmatrix}.$$

Next consider the eigenvectors for $\lambda = 1$. The matrix which must be row reduced is

$$\begin{pmatrix} \frac{13}{5} - 1 & \frac{2}{5} \sqrt{5} & \frac{8}{5} \sqrt{5} & | & 0 \\ \frac{2}{5} \sqrt{5} & \frac{4}{5} - 1 & \frac{4}{15} & | & 0 \\ \frac{8}{15} \sqrt{5} & \frac{4}{15} & \frac{61}{15} - 1 & | & 0 \end{pmatrix}$$

and its row reduced echelon form is

$$\begin{pmatrix} 1 & \frac{3}{10} \sqrt{5} & \frac{2}{5} \sqrt{5} & | & 0 \\ 0 & 0 & 0 & | & 0 \\ 0 & 0 & 0 & | & 0 \end{pmatrix}.$$ 

Therefore, the eigenvectors are of the form

$$\begin{pmatrix} -\frac{3}{10} \sqrt{5} y - \frac{2}{5} \sqrt{5} z \\ y \\ z \end{pmatrix}.$$ 

This is a two dimensional eigenspace.

Before going further, we want to point out that no matter how we choose $y$ and $z$ the resulting vector will be orthogonal to the eigenvector for $\lambda = 2$. This is a special case of a general result which states that eigenvectors for distinct eigenvalues of a symmetric matrix are orthogonal. This is explained in Problem 22. For this case you need to show the following dot product equals zero.

$$\begin{pmatrix} \frac{2}{3} \\ \frac{1}{5} \sqrt{5} \\ \frac{4}{15} \sqrt{5} \end{pmatrix} \cdot \begin{pmatrix} -\frac{3}{10} \sqrt{5} y - \frac{2}{5} \sqrt{5} z \\ y \\ z \end{pmatrix}$$

This is left for you to do.

Continuing with the task of finding an orthonormal basis, Let $y = 0$ first. This results in eigenvectors of the form

$$\begin{pmatrix} -\frac{2}{5} \sqrt{5} z \\ 0 \\ z \end{pmatrix}$$
and letting \( z = \frac{1}{3} \sqrt{5} \) you obtain a unit vector. Thus the second vector will be

\[
\begin{pmatrix}
-\frac{2}{3} \sqrt{5} \left( \frac{1}{3} \sqrt{5} \right) \\
0 \\
\frac{1}{3} \sqrt{5}
\end{pmatrix}
= \begin{pmatrix}
-\frac{2}{3} \\
0 \\
\frac{1}{3} \sqrt{5}
\end{pmatrix}.
\]

It remains to find the third vector in the orthonormal basis. This merely involves choosing \( y \) and \( z \) in (30.2) in such a way that the resulting vector has dot product with the two given vectors equal to zero. Thus you need

\[
\begin{pmatrix}
-\frac{2}{15} \sqrt{5} y - \frac{2}{5} \sqrt{5} z \\
y \\
z
\end{pmatrix}
\cdot
\begin{pmatrix}
-\frac{2}{3} \\
0 \\
\frac{1}{3} \sqrt{5}
\end{pmatrix}
= \frac{1}{5} \sqrt{5} y + \frac{3}{5} \sqrt{5} z = 0.
\]

The dot product with the eigenvector for \( \lambda = 2 \) is automatically equal to zero and so all that you need is the above equation. This is satisfied when \( z = -\frac{1}{3} y \). Therefore, the vector we want is of the form

\[
\begin{pmatrix}
-\frac{2}{15} \sqrt{5} y - \frac{2}{5} \sqrt{5} (-\frac{1}{3} y) \\
y \\
(-\frac{1}{3} y)
\end{pmatrix}
= \begin{pmatrix}
-\frac{1}{3} \sqrt{5} y \\
y \\
-\frac{1}{3} y
\end{pmatrix}.
\]

and it only remains to choose \( y \) in such a way that this vector has unit length. This occurs when \( y = \frac{2}{5} \sqrt{5} \). Therefore, the vector we want is

\[
\begin{pmatrix}
2 \sqrt{5} \\
-\frac{1}{3} \sqrt{5} \\
-\frac{2}{15} \sqrt{5}
\end{pmatrix}
= \begin{pmatrix}
-\frac{1}{3} \sqrt{5} \\
\frac{1}{3} \\
-\frac{2}{15} \sqrt{5}
\end{pmatrix}.
\]

The three eigenvectors which constitute an orthonormal basis are

\[
\begin{pmatrix}
-\frac{1}{3} \\
\frac{2}{5} \sqrt{5} \\
-\frac{2}{15} \sqrt{5}
\end{pmatrix},
\begin{pmatrix}
-\frac{2}{3} \\
0 \\
\frac{1}{3} \sqrt{5}
\end{pmatrix}, \text{ and }
\begin{pmatrix}
\frac{2}{5} \\
\frac{1}{5} \sqrt{5} \\
\frac{4}{15} \sqrt{5}
\end{pmatrix}.
\]

To check our work and see if this is really an orthonormal set of vectors, we make them the columns of a matrix and see if the resulting matrix is orthogonal. The matrix is

\[
\begin{pmatrix}
-\frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \\
\frac{2}{5} \sqrt{5} & 0 & \frac{1}{5} \sqrt{5} \\
-\frac{2}{15} \sqrt{5} & \frac{1}{3} \sqrt{5} & \frac{4}{15} \sqrt{5}
\end{pmatrix}
\]

This matrix times its transpose equals

\[
\begin{pmatrix}
-\frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \\
\frac{2}{5} \sqrt{5} & 0 & \frac{1}{5} \sqrt{5} \\
-\frac{2}{15} \sqrt{5} & \frac{1}{3} \sqrt{5} & \frac{4}{15} \sqrt{5}
\end{pmatrix}
\begin{pmatrix}
-\frac{1}{3} & \frac{2}{5} \sqrt{5} & -\frac{2}{15} \sqrt{5} \\
\frac{2}{3} & 0 & \frac{1}{3} \sqrt{5} \\
\frac{2}{3} & \frac{1}{3} \sqrt{5} & \frac{4}{15} \sqrt{5}
\end{pmatrix}
= \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}.
\]
and so this is indeed an orthonormal basis.

Because of the repeated eigenvalue, there would have been many other orthonormal bases which could have been obtained. It was pretty arbitrary for to take \( y = 0 \) in the above argument. We could just as easily have taken \( z = 0 \) or even \( y = z = 1 \). Any such change would have resulted in a different orthonormal basis. Geometrically, what is happening is the eigenspace for \( \lambda = 1 \) was two dimensional. It can be visualized as a plane in three dimensional space which passes through the origin. There are infinitely many different pairs of perpendicular unit vectors in this plane.

### 30.1.3 Diagonalizing A Symmetric Matrix

Recall the following definition:

**Definition 30.1.17** An \( n \times n \) matrix, \( A = (a_{ij}) \) is called a diagonal matrix if \( a_{ij} = 0 \) whenever \( i \neq j \). For example, a diagonal matrix is of the form indicated below where \(*\) denotes a number.

\[
\begin{pmatrix}
* & 0 & \cdots & 0 \\
0 & * & \vdots \\
\vdots & \ddots & 0 \\
0 & \cdots & 0 & *
\end{pmatrix}
\]

As indicated in Theorem 30.1.13 if \( A \) is a real symmetric matrix, there exists an orthogonal matrix, \( U \) such that \( U^T A U = D \) a diagonal matrix. In the following example, this orthogonal matrix will be found.

**Example 30.1.18** Let \( A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{3}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{3}{2} \end{pmatrix} \). Find an orthogonal matrix, \( U \) such that \( U^T A U \) is a diagonal matrix.

In this case, a tedious computation shows the eigenvalues are 2 and 1. First we will find an eigenvector for the eigenvalue 2. This involves row reducing the following augmented matrix.

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

The row reduced echelon form is

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

and so an eigenvector is

\[
\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.
\]

However, it is desired that the eigenvectors obtained all be unit vectors and so dividing this vector by its length gives

\[
\begin{pmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}.
\]
Next consider the case of the eigenvalue, 1. The matrix which needs to be row reduced in this case is

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

The row reduced echelon form is

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

Therefore, the eigenvectors are of the form

\[
\begin{pmatrix}
s \\
-t \\
t \\
\end{pmatrix}
\]

Two of these which are orthonormal are

\[
\begin{pmatrix}
1 \\
0 \\
0 \\
\end{pmatrix}
\] and \[
\begin{pmatrix}
0 \\
-1/\sqrt{2} \\
1/\sqrt{2} \\
\end{pmatrix}
\]

An orthogonal matrix which works in the process is then obtained by letting these vectors be the columns.

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

It remains to verify this works. \(U^T AU\) is of the form

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

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

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

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

the desired diagonal matrix.

### 30.2 More General Considerations

All of the above topics generalize to complex matrices. You can read these in the appendix called Heroic Linear Algebra.

### 30.3 Exercises

1. Here are some matrices. Label according to whether they are symmetric, skew symmetric, or orthogonal. If the matrix is orthogonal, determine whether it is proper or improper.
 SOME SPECIAL MATRICES

(a) \[
\begin{pmatrix}
1 & 0 & 0 \\
0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\
0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}
\end{pmatrix}
\]

(b) \[
\begin{pmatrix}
1 & 2 & -3 \\
2 & 1 & 4 \\
-3 & 4 & 7
\end{pmatrix}
\]

(c) \[
\begin{pmatrix}
0 & -2 & -3 \\
2 & 0 & -4 \\
3 & 4 & 0
\end{pmatrix}
\]

2. Here are some matrices. What can you say about the eigenvalues of these matrices just by looking at them?

(a) \[
\begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & -1 \\
0 & 1 & 0
\end{pmatrix}
\]

(b) \[
\begin{pmatrix}
1 & 2 & -3 \\
2 & 1 & 4 \\
-3 & 4 & 7
\end{pmatrix}
\]

(c) \[
\begin{pmatrix}
0 & -2 & -3 \\
2 & 0 & -4 \\
3 & 4 & 0
\end{pmatrix}
\]

(d) \[
\begin{pmatrix}
1 & 2 & 3 \\
0 & 2 & 3 \\
0 & 0 & 2
\end{pmatrix}
\]

3. Explain why a matrix, \( A \) is symmetric if and only if there exists an orthogonal matrix, \( U \) such that \( A = U^T D U \) for \( D \) a diagonal matrix.

4. If \( A \) is a symmetric invertible matrix, is it always the case that \( A^{-1} \) must be symmetric also? How about \( A^k \) for \( k \) a positive integer? Explain.

5. For \( U \) an orthogonal matrix, explain why \( \|Ux\| = \|x\| \) for any vector, \( x \). Next explain why if \( U \) is an \( n \times n \) matrix with the property that \( \|Ux\| = \|x\| \) for all vectors, \( x \), then \( U \) must be orthogonal. Thus the orthogonal matrices are exactly those which preserve distance.

6. Let \( x \) be a vector in \( \mathbb{R}^n \) and consider the matrix, \( I - \frac{2xx^T}{\|x\|^2} \). Show this matrix is both symmetric and orthogonal.

7. A quadratic form in three variables is an expression of the form \( a_1x^2 + a_2y^2 + a_3z^2 + a_4xy + a_5xz + a_6yz \). Show that every such quadratic form may be written as

\[
\begin{pmatrix}
x & y & z
\end{pmatrix}
A
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
\]

where \( A \) is a symmetric matrix.
8. Given a quadratic form in three variables, \( x, y, \) and \( z, \) show there exists an orthogonal matrix, \( U \) and variables \( x', y', z' \) such that

\[
\begin{pmatrix}
x \\
y \\
z \\
\end{pmatrix} = U \begin{pmatrix}
x' \\
y' \\
z' \\
\end{pmatrix}
\]

with the property that in terms of the new variables, the quadratic form is

\[
\lambda_1 (x')^2 + \lambda_2 (y')^2 + \lambda_3 (z')^2
\]

where the numbers, \( \lambda_1, \lambda_2, \) and \( \lambda_3 \) are the eigenvalues of the matrix, \( A \) in Problem 7.

9. Suppose \( A \) is a \( 3 \times 3 \) symmetric matrix and you have found two eigenvectors which form an orthonormal set. Explain why their cross product is also an eigenvector.

10. Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix,

\[
\begin{pmatrix}
11 & -1 & -4 \\
-1 & 11 & -4 \\
-4 & -4 & 14 \\
\end{pmatrix}
\]

**Hint:** Two eigenvalues are 12 and 18.

11. Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix,

\[
\begin{pmatrix}
4 & 1 & -2 \\
1 & 4 & -2 \\
-2 & -2 & 7 \\
\end{pmatrix}
\]

**Hint:** One eigenvalue is 3.

12. Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix,

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

**Hint:** One eigenvalue is -2.

13. Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix,

\[
\begin{pmatrix}
17 & -7 & -4 \\
-7 & 17 & -4 \\
-4 & -4 & 14 \\
\end{pmatrix}
\]

**Hint:** Two eigenvalues are 18 and 24.

14. Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix,

\[
\begin{pmatrix}
13 & 1 & 4 \\
1 & 13 & 4 \\
4 & 4 & 10 \\
\end{pmatrix}
\]

**Hint:** Two eigenvalues are 12 and 18.
15. Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix,

\[
\begin{pmatrix}
-\frac{5}{3} & \frac{1}{15}\sqrt{6}\sqrt{5} & \frac{8}{15}\sqrt{5} \\
\frac{1}{15}\sqrt{6}\sqrt{5} & -\frac{14}{5} & -\frac{1}{15}\sqrt{6} \\
\frac{8}{15}\sqrt{5} & -\frac{1}{15}\sqrt{6} & \frac{2}{5}
\end{pmatrix}
\]

**Hint:** The eigenvalues are $-3, -2, 1$.

16. Let \( A = \)

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

Find an orthogonal matrix, \( U \) such that \( U^T A U \) is a diagonal matrix.

17. Let \( A = \)

\[
\begin{pmatrix}
2 & 0 & 0 \\
0 & 5 & 1 \\
0 & 1 & 5
\end{pmatrix}
\]

Find an orthogonal matrix, \( U \) such that \( U^T A U \) is a diagonal matrix.

18. Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix,

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

**Hint:** The eigenvalues are $0, 2, 2$ where $2$ is listed twice because it is a root of multiplicity 2.

19. Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix,

\[
\begin{pmatrix}
1 & \frac{1}{6}\sqrt{3}\sqrt{2} & \frac{1}{6}\sqrt{3}\sqrt{6} \\
\frac{1}{6}\sqrt{3}\sqrt{2} & \frac{3}{2} & \frac{1}{12}\sqrt{2}\sqrt{6} \\
\frac{1}{6}\sqrt{3}\sqrt{6} & \frac{1}{12}\sqrt{2}\sqrt{6} & \frac{1}{2}
\end{pmatrix}
\]

**Hint:** The eigenvalues are $2, 1, 0$.

20. Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix,
30.3. EXERCISES 547

\[
\begin{pmatrix}
\frac{1}{3} & \frac{1}{6} \sqrt{3} \sqrt{2} & -\frac{7}{18} \sqrt{3} \sqrt{6} \\
\frac{1}{6} \sqrt{3} \sqrt{2} & \frac{1}{2} & -\frac{1}{12} \sqrt{2} \sqrt{6} \\
-\frac{7}{18} \sqrt{3} \sqrt{6} & -\frac{1}{12} \sqrt{2} \sqrt{6} & -\frac{5}{6}
\end{pmatrix}
\]

**Hint:** The eigenvalues are 1, 2, −2.

21. Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix,

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

**Hint:** The eigenvalues are −1, 2, −1 where −1 is listed twice because it has multiplicity 2 as a zero of the characteristic equation.

22. Show that if \(A\) is a real symmetric matrix and \(\lambda\) and \(\mu\) are two different eigenvalues, then if \(x\) is an eigenvector for \(\lambda\) and \(y\) is an eigenvector for \(\mu\), then \(x \cdot y = 0\). Also all eigenvalues are real. Supply reasons for each step in the following argument. First

\[
\lambda x^T x = (Ax)^T x = x^T Ax = x^T \overline{Ax} = \overline{x}^T \overline{Ax} = \overline{\lambda} x^T \overline{x}
\]

and so \(\lambda = \overline{\lambda}\). This shows that all eigenvalues are real. It follows all the eigenvectors are real. Why? Now let \(x, y, \mu\) and \(\lambda\) be given as above.

\[
\lambda (x \cdot y) = \lambda x \cdot y = Ax \cdot y = x \cdot Ay = x \cdot \mu y = \mu (x \cdot y) = \mu (x \cdot y)
\]

and so

\[
(\lambda - \mu) x \cdot y = 0.
\]

Since \(\lambda \neq \mu\), it follows \(x \cdot y = 0\).

23. If a real matrix, \(A\) has all real eigenvalues, does it follow that \(A\) must be symmetric. If so, explain why and if not, give an example to the contrary.

24. Fill in the missing entries to make the matrix orthogonal.

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

25. Show that if \(\{u_1, \cdots, u_n\}\) is an orthonormal set of vectors in \(\mathbb{F}^n\), then it is a basis.

**Hint:** It was shown earlier that this is a linearly independent set. Now review the exchange theorem and verify that because of this, it must be a basis. If you wish, replace \(\mathbb{F}^n\) with \(\mathbb{R}^n\). Do this version if you do not know the dot product for vectors in \(\mathbb{C}^n\).
26. Suppose $U$ is an orthogonal $n \times n$ matrix. Explain why $\text{rank}(U) = n$.

27. Show that if $A$ is an Hermitian matrix and $\lambda$ and $\mu$ are two different eigenvalues, then if $x$ is an eigenvector for $\lambda$ and $y$ is an eigenvector for $\mu$, then $x \cdot y = 0$. Also all eigenvalues are real. Supply reasons for each step in the following argument. First

\[ \lambda x \cdot x = \lambda x = A x \cdot x = x \cdot \lambda x = \overline{\lambda} x \cdot x \]

and so $\lambda = \overline{\lambda}$. This shows that all eigenvalues are real. Now let $x, y, \mu$ and $\lambda$ be given as above.

\[ \lambda (x \cdot y) = \lambda x \cdot y = A x \cdot y = x \cdot \lambda y = \overline{\mu} (x \cdot y) = \mu (x \cdot y) \]

and so

\[ (\lambda - \mu) x \cdot y = 0. \]

Since $\lambda \neq \mu$, it follows $x \cdot y = 0$.

28. Show that the eigenvalues and eigenvectors of a real matrix occur in conjugate pairs.

29. Show that every real matrix may be written as the sum of a skew symmetric and a symmetric matrix. Hint: If $A$ is an $n \times n$ matrix, show that $B \equiv \frac{1}{2} (A - A^T)$ is skew symmetric.

30. The set of real $n \times n$ matrices forms a vector space. Explain why. Explain why the set of $n \times n$ symmetric matrices forms a subspace. What is the dimension of this subspace and why?

31. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
c & 0 & 0 \\
0 & 0 & -b \\
0 & b & 0
\end{pmatrix}
\]

Here $b, c$ are real numbers.

32. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
c & 0 & 0 \\
0 & a & -b \\
0 & b & a
\end{pmatrix}
\]

Here $a, b, c$ are real numbers.
Numerical Methods For Solving Linear Systems

31.0.1 Outcomes

1. Apply Gauss-Seidel iteration to approximate a solution to a linear system of equations.
2. Apply Jacobi iteration to approximate a solution to a linear system of equations.

31.1 Iterative Methods For Linear Systems

Consider the problem of solving the equation

\[ Ax = b \]  

(31.1)

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. The idea is to obtain a sequence of approximate solutions which get close to the true solution after a sufficient number of iterations.

**Definition 31.1.1** Let \( \{x_k\}_{k=1}^{\infty} \) be a sequence of vectors in \( \mathbb{F}^n \). Say

\[ x_k = (x_{k,1}, \cdots, x_{k,n}) \]

Then this sequence is said to converge to the vector, \( x = (x_1, \cdots, x_n) \in \mathbb{F}^n \), written as

\[ \lim_{k \to \infty} x_k = x \]

if for each \( j = 1, 2, \cdots, n \),

\[ \lim_{k \to \infty} x_{k,j} = x_j. \]

In words, the sequence converges if the entries of the vectors in the sequence converge to the corresponding entries of \( x \).

**Example 31.1.2** Consider \( x_k = \left( \sin\left(\frac{1}{k}\right), \frac{k^2}{1+k^2}, \ln\left(\frac{1+k^2}{k^2}\right) \right) \). Find \( \lim_{k \to \infty} x_k \).

From the above definition, this limit is the vector, \((0, 1, 0)\) because

\[ \lim_{k \to \infty} \sin\left(\frac{1}{k}\right) = 0, \quad \lim_{k \to \infty} \frac{k^2}{1+k^2} = 1, \quad \text{and} \quad \lim_{k \to \infty} \ln\left(\frac{1+k^2}{k^2}\right) = 0. \]

See the appendix called Heroic Linear Algebra.
31.1.1 The Jacobi Method

The first technique to be discussed here is the Jacobi method which is described in the following definition. In this technique, you have a sequence of vectors, \( \{ \mathbf{x}^k \} \) which converge to the solution to the linear system of equations and to get the \( i \)th component of the \( \mathbf{x}^{k+1} \), you use all the components of \( \mathbf{x}^k \) except for the \( i \)th. The precise description follows.

Definition 31.1.3 The Jacobi iterative technique, also called the method of simultaneous corrections, is defined as follows. Let \( \mathbf{x}^1 \) be an initial vector, say the zero vector or some other vector. The method generates a succession of vectors, \( \mathbf{x}^2, \mathbf{x}^3, \mathbf{x}^4, \cdots \) and hopefully this sequence of vectors will converge to the solution to (31.1). The vectors in this list are called iterates and they are obtained according to the following procedure. Letting \( A = (a_{ij}) \),

\[
a_{ii}x_{r+1}^i = - \sum_{j \neq i} a_{ij} x_r^j + b_i. \tag{31.2}
\]

In terms of matrices, letting

\[
A = \begin{pmatrix}
a_{11} & \cdots & a_{1n} \\
\vdots & \ddots & \vdots \\
a_{n1} & \cdots & a_{nn}
\end{pmatrix}
\]

The iterates are defined as

\[
\begin{pmatrix}
a_{11} & 0 & \cdots & 0 \\
0 & a_{22} & \cdots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & a_{nn}
\end{pmatrix}
\begin{pmatrix}
x_1^{r+1} \\
x_2^{r+1} \\
\vdots \\
x_n^{r+1}
\end{pmatrix}
= - \begin{pmatrix}
0 & a_{12} & \cdots & a_{1n} \\
a_{21} & 0 & \cdots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
a_{n1} & \cdots & a_{nn-1} & 0
\end{pmatrix}
\begin{pmatrix}
x_1^r \\
x_2^r \\
\vdots \\
x_n^r
\end{pmatrix}
+ \begin{pmatrix}
b_1 \\
b_2 \\
\vdots \\
b_n
\end{pmatrix} \tag{31.3}
\]

The matrix on the left in (31.3) is obtained by retaining the main diagonal of \( A \) and setting every other entry equal to zero. The matrix on the right in (31.3) is obtained from \( A \) by setting every diagonal entry equal to zero and retaining all the other entries unchanged.

Example 31.1.4 Use the Jacobi method to solve the system

\[
\begin{pmatrix}
3 & 1 & 0 & 0 \\
1 & 4 & 1 & 0 \\
0 & 2 & 5 & 1 \\
0 & 0 & 2 & 4
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{pmatrix}
= \begin{pmatrix}
1 \\
2 \\
3 \\
4
\end{pmatrix}
\]

In terms of the matrices, the Jacobi iteration is of the form

\[
\begin{pmatrix}
3 & 0 & 0 & 0 \\
0 & 4 & 0 & 0 \\
0 & 0 & 5 & 0 \\
0 & 0 & 0 & 4
\end{pmatrix}
\begin{pmatrix}
x_1^{r+1} \\
x_2^{r+1} \\
x_3^{r+1} \\
x_4^{r+1}
\end{pmatrix}
= - \begin{pmatrix}
0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
0 & 2 & 0 & 1 \\
0 & 0 & 2 & 0
\end{pmatrix}
\begin{pmatrix}
x_1^r \\
x_2^r \\
x_3^r \\
x_4^r
\end{pmatrix}
+ \begin{pmatrix}
1 \\
2 \\
3 \\
4
\end{pmatrix}
\]

Now iterate this starting with
31.1. Iterative Methods for Linear Systems

Let $x^1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$.

\[
\begin{pmatrix}
3 & 0 & 0 & 0 \\
0 & 4 & 0 & 0 \\
0 & 0 & 5 & 0 \\
0 & 0 & 0 & 4
\end{pmatrix}
\begin{pmatrix}
x_1^2 \\
x_2^2 \\
x_3^2 \\
x_4^2
\end{pmatrix} = -\begin{pmatrix}
0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
0 & 2 & 0 & 1 \\
0 & 0 & 2 & 0
\end{pmatrix}
\begin{pmatrix}
x_1^3 \\
x_2^3 \\
x_3^3 \\
x_4^3
\end{pmatrix} + \begin{pmatrix}
1 \\
2 \\
3 \\
4
\end{pmatrix}
\]

\[
\begin{pmatrix}
x_1^2 \\
x_2^2 \\
x_3^2 \\
x_4^2
\end{pmatrix} = \begin{pmatrix}
1.0 \\
2.0 \\
3.0 \\
4.0
\end{pmatrix}
\]

Solving this system yields

\[
x^2 = \begin{pmatrix}
x_1^2 \\
x_2^2 \\
x_3^2 \\
x_4^2
\end{pmatrix} = \begin{pmatrix}
0.333333 \\
0.5 \\
0.6 \\
1.0
\end{pmatrix}
\]

Then you use $x^2$ to find $x^3 = \begin{pmatrix} x_1^3 \\ x_2^3 \\ x_3^3 \\ x_4^3 \end{pmatrix}^T$

\[
\begin{pmatrix}
3 & 0 & 0 & 0 \\
0 & 4 & 0 & 0 \\
0 & 0 & 5 & 0 \\
0 & 0 & 0 & 4
\end{pmatrix}
\begin{pmatrix}
x_1^3 \\
x_2^3 \\
x_3^3 \\
x_4^3
\end{pmatrix} = -\begin{pmatrix}
0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
0 & 2 & 0 & 1 \\
0 & 0 & 2 & 0
\end{pmatrix}
\begin{pmatrix}
x_1^4 \\
x_2^4 \\
x_3^4 \\
x_4^4
\end{pmatrix} + \begin{pmatrix}
.333333 \\
.5 \\
.6 \\
1.0
\end{pmatrix}
\]

\[
\begin{pmatrix}
x_1^3 \\
x_2^3 \\
x_3^3 \\
x_4^3
\end{pmatrix} = \begin{pmatrix}
1.066667 \\
1.0 \\
2.8
\end{pmatrix}
\]

The solution is

\[
x^3 = \begin{pmatrix}
x_1^3 \\
x_2^3 \\
x_3^3 \\
x_4^3
\end{pmatrix} = \begin{pmatrix}
.166666 \\
.26666 \\
.2 \\
.7
\end{pmatrix}
\]

Now use this as the new data to find $x^4 = \begin{pmatrix} x_1^4 \\ x_2^4 \\ x_3^4 \\ x_4^4 \end{pmatrix}^T$

\[
\begin{pmatrix}
3 & 0 & 0 & 0 \\
0 & 4 & 0 & 0 \\
0 & 0 & 5 & 0 \\
0 & 0 & 0 & 4
\end{pmatrix}
\begin{pmatrix}
x_1^4 \\
x_2^4 \\
x_3^4 \\
x_4^4
\end{pmatrix} = -\begin{pmatrix}
0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
0 & 2 & 0 & 1 \\
0 & 0 & 2 & 0
\end{pmatrix}
\begin{pmatrix}
x_1^5 \\
x_2^5 \\
x_3^5 \\
x_4^5
\end{pmatrix} + \begin{pmatrix}
.166666 \\
.26666 \\
.2 \\
.7
\end{pmatrix}
\]

\[
\begin{pmatrix}
x_1^4 \\
x_2^4 \\
x_3^4 \\
x_4^4
\end{pmatrix} = \begin{pmatrix}
.733333 \\
1.633333 \\
1.766666 \\
3.6
\end{pmatrix}
\]

Thus you find

\[
x^4 = \begin{pmatrix}
.244444 \\
.408333 \\
.353333 \\
.9
\end{pmatrix}
\]
Then another iteration for $x^5$ gives

$$
\begin{pmatrix}
3 & 0 & 0 & 0 \\
0 & 4 & 0 & 0 \\
0 & 0 & 5 & 0 \\
0 & 0 & 0 & 4
\end{pmatrix}
\begin{pmatrix}
x_1^5 \\
x_2^5 \\
x_3^5 \\
x_4^5
\end{pmatrix}
= - \begin{pmatrix}
0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
0 & 2 & 0 & 1 \\
0 & 0 & 2 & 0
\end{pmatrix}
\begin{pmatrix}
.24444444 \\
.40833333 \\
.35333332 \\
.9
\end{pmatrix}
+ \begin{pmatrix}
1 \\
2 \\
3 \\
4
\end{pmatrix}
$$

$$
= \begin{pmatrix}
.59166667 \\
1.40222222 \\
1.28333333 \\
3.29333344
\end{pmatrix}
$$

and so

$$
x^5 = \begin{pmatrix}
.19722222 \\
.35055555 \\
.25666666 \\
.82333335
\end{pmatrix}
$$

The solution to the system of equations obtained by row operations is

$$
\begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{pmatrix}
= \begin{pmatrix}
.206 \\
.379 \\
.275 \\
.862
\end{pmatrix}
$$

so already after only five iterations the iterates are pretty close to the true solution. How well does it work?

$$
\begin{pmatrix}
3 & 1 & 0 & 0 \\
1 & 4 & 1 & 0 \\
0 & 2 & 5 & 1 \\
0 & 0 & 2 & 4
\end{pmatrix}
\begin{pmatrix}
x_1^5 \\
x_2^5 \\
x_3^5 \\
x_4^5
\end{pmatrix}
= - \begin{pmatrix}
0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
0 & 2 & 0 & 1 \\
0 & 0 & 2 & 0
\end{pmatrix}
\begin{pmatrix}
.24444444 \\
.40833333 \\
.35333332 \\
.9
\end{pmatrix}
+ \begin{pmatrix}
1 \\
2 \\
3 \\
4
\end{pmatrix}
$$

$$
\approx \begin{pmatrix}
1 \\
2 \\
3 \\
4
\end{pmatrix}
$$

A few more iterates will yield a better solution.

### 31.1.2 The Gauss Seidel Method

The Gauss Seidel method differs from the Jacobi method in using $x_{j}^{k+1}$ for all $j < i$ in going from $x^k$ to $x^{k+1}$. This is why it is called the method of successive corrections. The precise description of this method is in the following definition.

**Definition 31.1.5** 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

$$
\sum_{j=1}^{i} a_{ij} x_j^{r+1} = - \sum_{j=i+1}^{n} a_{ij} x_j^r + b_i.
$$

(31.4)

In terms of matrices, letting

$$
A = \begin{pmatrix}
a_{11} & \cdots & a_{1n} \\
\vdots & \ddots & \vdots \\
a_{n1} & \cdots & a_{nn}
\end{pmatrix}
$$
The iterates are defined as

\[
\begin{pmatrix}
  a_{11} & 0 & \cdots & 0 \\
  a_{21} & a_{22} & \ddots & \vdots \\
  \vdots & \ddots & \ddots & 0 \\
  a_{n1} & \cdots & a_{n-1,n} & a_{nn}
\end{pmatrix}
\begin{pmatrix}
  x_{r+1}^1 \\
  x_{r+1}^2 \\
  \vdots \\
  x_{r+1}^n
\end{pmatrix}
= -
\begin{pmatrix}
  0 & a_{12} & \cdots & a_{1n} \\
  0 & 0 & \ddots & \vdots \\
  \vdots & \ddots & \ddots & a_{n-1,n} \\
  0 & \cdots & 0 & 0
\end{pmatrix}
\begin{pmatrix}
  x_r^1 \\
  x_r^2 \\
  \vdots \\
  x_r^n
\end{pmatrix}
+ 
\begin{pmatrix}
  b_1 \\
  b_2 \\
  \vdots \\
  b_n
\end{pmatrix}
\tag{31.5}
\]

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 of (31.4) 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 31.1.6** 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

\[
\begin{pmatrix}
  3 & 0 & 0 & 0 \\
  1 & 4 & 0 & 0 \\
  0 & 2 & 5 & 0 \\
  0 & 0 & 2 & 4
\end{pmatrix}
\begin{pmatrix}
  x_{r+1}^1 \\
  x_{r+1}^2 \\
  x_{r+1}^3 \\
  x_{r+1}^4
\end{pmatrix}
= -
\begin{pmatrix}
  0 & 1 & 0 & 0 \\
  0 & 0 & 1 & 0 \\
  0 & 0 & 0 & 1 \\
  0 & 0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
  x_r^1 \\
  x_r^2 \\
  x_r^3 \\
  x_r^4
\end{pmatrix}
+ 
\begin{pmatrix}
  1 \\
  2 \\
  3 \\
  4
\end{pmatrix}
\]

As before, let \( x^1 \) be the zero vector. Thus the first iteration is to solve

\[
\begin{pmatrix}
  3 & 0 & 0 & 0 \\
  1 & 4 & 0 & 0 \\
  0 & 2 & 5 & 0 \\
  0 & 0 & 2 & 4
\end{pmatrix}
\begin{pmatrix}
  x_1^2 \\
  x_2^2 \\
  x_3^2 \\
  x_4^2
\end{pmatrix}
= -
\begin{pmatrix}
  0 & 1 & 0 & 0 \\
  0 & 0 & 1 & 0 \\
  0 & 0 & 0 & 1 \\
  0 & 0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
  x_1^1 \\
  x_2^1 \\
  x_3^1 \\
  x_4^1
\end{pmatrix}
+ 
\begin{pmatrix}
  1 \\
  2 \\
  3 \\
  4
\end{pmatrix}
\]

Hence

\[
x^2 = \begin{pmatrix}
  0.33333333 \\
  0.41666667 \\
  0.43333333 \\
  0.78333333
\end{pmatrix}
\]
Thus $x^3 = (x_1^3 \ x_2^3 \ x_3^3 \ x_4^3)^T$ is given by
\[
\begin{pmatrix}
3 & 0 & 0 & 0 \\
1 & 4 & 0 & 0 \\
0 & 2 & 5 & 0 \\
0 & 0 & 2 & 4
\end{pmatrix}
\begin{pmatrix}
x_1^3 \\
x_2^3 \\
x_3^3 \\
x_4^3
\end{pmatrix}
= -
\begin{pmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
.33333333 \\
.41666667 \\
.43333333 \\
.78333333
\end{pmatrix}
+ 
\begin{pmatrix}
1 \\
2 \\
3 \\
4
\end{pmatrix}
= 
\begin{pmatrix}
.58333333 \\
1.56666667 \\
2.21666667 \\
4.0
\end{pmatrix}
\]
And so
\[
x^3 = 
\begin{pmatrix}
.19444444 \\
.34305556 \\
.30611111 \\
.84694444
\end{pmatrix}
\]
Another iteration for $x^4$ involves solving
\[
\begin{pmatrix}
3 & 0 & 0 & 0 \\
1 & 4 & 0 & 0 \\
0 & 2 & 5 & 0 \\
0 & 0 & 2 & 4
\end{pmatrix}
\begin{pmatrix}
x_1^4 \\
x_2^4 \\
x_3^4 \\
x_4^4
\end{pmatrix}
= -
\begin{pmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
.19444444 \\
.34305556 \\
.30611111 \\
.84694444
\end{pmatrix}
+ 
\begin{pmatrix}
1 \\
2 \\
3 \\
4
\end{pmatrix}
= 
\begin{pmatrix}
.65694444 \\
1.69388889 \\
2.15305556 \\
4.0
\end{pmatrix}
\]
and so
\[
x^4 = 
\begin{pmatrix}
.21898148 \\
.36872686 \\
.28312038 \\
.85843981
\end{pmatrix}
\]
Recall the answer is
\[
\begin{pmatrix}
.206 \\
.379 \\
.275 \\
.862
\end{pmatrix}
\]
so the iterates are already pretty close to the answer. You could continue doing these iterates and it appears they converge to the solution. Now consider the following example.

**Example 31.1.7** *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 row echelon form is
\[
\begin{pmatrix}
1 & 0 & 0 & 0 & 6 \\
0 & 1 & 0 & 0 & -\frac{5}{4} \\
0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & \frac{1}{2}
\end{pmatrix}
\]
and so the solution is approximately

\[
\begin{pmatrix}
\frac{6}{5} \\
\frac{1}{2}
\end{pmatrix} =
\begin{pmatrix}
6.0 \\ -1.25
\end{pmatrix}
\]

The Gauss Seidel iterations are of the form

\[
\begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 4 & 0 & 0 \\
0 & 2 & 5 & 0 \\
0 & 0 & 2 & 4
\end{pmatrix}
\begin{pmatrix}
x^{r+1}_1 \\
x^{r+1}_2 \\
x^{r+1}_3 \\
x^{r+1}_4
\end{pmatrix} =
\begin{pmatrix}
0 & 4 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
x^{r}_1 \\
x^{r}_2 \\
x^{r}_3 \\
x^{r}_4
\end{pmatrix} +
\begin{pmatrix}
1 \\
2 \\
3 \\
4
\end{pmatrix}
\]

and so, multiplying by the inverse of the matrix on the left, the iteration reduces to the following in terms of matrix multiplication.

\[
x^{r+1} = -
\begin{pmatrix}
0 & 4 & 0 & 0 \\
0 & -1 & \frac{1}{4} & 0 \\
0 & \frac{2}{5} & -\frac{1}{10} & \frac{1}{5} \\
0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10}
\end{pmatrix}
\begin{pmatrix}
x^{r}_1 \\
x^{r}_2 \\
x^{r}_3 \\
x^{r}_4
\end{pmatrix} +
\begin{pmatrix}
1 \\
\frac{1}{4} \\
\frac{1}{2} \\
\frac{3}{4}
\end{pmatrix}
\]

This time, we will pick an initial vector close to the answer. Let

\[
x^1 = \begin{pmatrix}
6 \\
-1 \\
1 \\
\frac{1}{2}
\end{pmatrix}
\]

This is very close to the answer. Now lets see what the Gauss Seidel iteration does to it.

\[
x^2 = -
\begin{pmatrix}
0 & 4 & 0 & 0 \\
0 & -1 & \frac{1}{4} & 0 \\
0 & \frac{2}{5} & -\frac{1}{10} & \frac{1}{5} \\
0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10}
\end{pmatrix}
\begin{pmatrix}
6 \\
-1 \\
1 \\
\frac{3}{4}
\end{pmatrix} +
\begin{pmatrix}
1 \\
\frac{1}{4} \\
\frac{1}{2} \\
\frac{3}{4}
\end{pmatrix} =
\begin{pmatrix}
5.0 \\
-1.0 \\
.9 \\
.55
\end{pmatrix}
\]

You can’t expect to be real close after only one iteration. Lets do another.

\[
x^3 = -
\begin{pmatrix}
0 & 4 & 0 & 0 \\
0 & -1 & \frac{1}{4} & 0 \\
0 & \frac{2}{5} & -\frac{1}{10} & \frac{1}{5} \\
0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10}
\end{pmatrix}
\begin{pmatrix}
5.0 \\
-1.0 \\
.9 \\
.55
\end{pmatrix} +
\begin{pmatrix}
1 \\
\frac{1}{4} \\
\frac{1}{2} \\
\frac{3}{4}
\end{pmatrix} =
\begin{pmatrix}
5.0 \\
-1.975 \\
.88 \\
.56
\end{pmatrix}
\]

\[
x^4 = -
\begin{pmatrix}
0 & 4 & 0 & 0 \\
0 & -1 & \frac{1}{4} & 0 \\
0 & \frac{2}{5} & -\frac{1}{10} & \frac{1}{5} \\
0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10}
\end{pmatrix}
\begin{pmatrix}
5.0 \\
-1.975 \\
.88 \\
.56
\end{pmatrix} +
\begin{pmatrix}
1 \\
\frac{1}{4} \\
\frac{1}{2} \\
\frac{3}{4}
\end{pmatrix} =
\begin{pmatrix}
4.9 \\
-.945 \\
.866 \\
.567
\end{pmatrix}
\]
The iterates seem to be getting farther from 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? A complete answer to this question is given in the appendix called Heroic Linear algebra but here we will simply state a condition which guarantees the process will work.

Both iterative procedures for solving

$$Ax = b$$

are of the form

$$Bx^{r+1} = -Cx^r + b$$

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

$$x^{r+1} = -B^{-1}Cx^r + B^{-1}b.$$ 

(31.7)

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.

**Theorem 31.1.8** Let $A = B + C$ and suppose all eigenvalues of $B^{-1}C$ have absolute value less than 1 where $A = B + C$. Then the iterates in (31.7) converge to the unique solution of (31.6).

A complete explanation of this important result is found in the appendix, Heroic Linear Algebra. It depends on a theorem of Gelfand which is completely proved in this appendix. Theorem 31.1.8 is very remarkable because it gives an algebraic condition for convergence which is essentially an analytical question.

### 31.2 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 using either of these methods for finding eigenvectors for a given eigenvalue?
Numerical Methods For Solving The Eigenvalue Problem

32.0.1 Outcomes

1. Apply the power method with scaling to approximate the dominant eigenvector corresponding to a dominant eigenvalue.

2. Use the shifted inverse power method to find the eigenvector and eigenvalue close to some number.

3. Approximate an eigenvalue of a symmetric matrix by computing Rayleigh quotient and finding the associated error bound. Illustrate why the Rayleigh quotient approximates the dominant eigenvalue.

32.1 The Power Method For Eigenvalues

As indicated earlier, the eigenvalue eigenvector problem is extremely difficult. Consider for example what happens if you cannot find the eigenvalues exactly. Then you can’t find an eigenvector because there isn’t one due to the fact that \( A - \lambda I \) is invertible whenever \( \lambda \) is not exactly equal to an eigenvalue. Therefore the straightforward way of solving this problem fails right away, even if you can approximate the eigenvalues. The power method allows you to approximate the largest eigenvalue and also the eigenvector which goes with it. By considering the inverse of the matrix, you can also find the smallest eigenvalue. The method works in the situation of a nondefective matrix, \( A \) which has an eigenvalue of algebraic multiplicity 1, \( \lambda_n \) which has the property that \( |\lambda_k| < |\lambda_n| \) for all \( k \neq n \). Note that for a real matrix this excludes the case that \( \lambda_n \) could be complex. Why? Such an eigenvalue is called a dominant eigenvalue.

Let \( \{x_1, \ldots, x_n\} \) be a basis of eigenvectors for \( \mathbb{F}^n \) such that \( Ax_n = \lambda_n x_n \). Now let \( u_1 \) be some nonzero vector. Since \( \{x_1, \ldots, x_n\} \) is a basis, there exists unique scalars, \( c_i \) such that

\[
 u_1 = \sum_{k=1}^{n} c_k x_k.
\]

Assume you have not been so unlucky as to pick \( u_1 \) in such a way that \( c_n = 0 \). Then let \( Au_k = u_{k+1} \) so that

\[
 u_m = A^m u_1 = \sum_{k=1}^{n-1} c_k \lambda_k^m x_k + \lambda_n^m c_n x_n. \tag{32.1}
\]

559
For large \( m \) the last term, \( \lambda_n^m c_n x_n \), determines quite well the direction of the vector on the right. This is because \(|\lambda_n|\) is larger than \(|\lambda_k|\) and so for a large, \( m \), the sum, \( \sum_{k=1}^{n-1} c_k \lambda_k^m x_k \), on the right is fairly insignificant. Therefore, for large \( m \), \( u_m \) is essentially a multiple of the eigenvector, \( x_n \), the one which goes with \( \lambda_n \). The only problem is that there is no control of the size of the vectors \( u_m \). You can fix this by scaling. Let \( S_2 \) denote the entry of \( Au_1 \) which is largest in absolute value. Then \( u_2 \) will not be just \( Au_1 \) but \( Au_1 / S_2 \). Next let \( S_3 \) denote the entry of \( Au_2 \) which has largest absolute value and define \( u_3 \equiv Au_2 / S_3 \). Continue this way. The scaling just described does not destroy the relative insignificance of the term involving a sum in (32.1). Indeed it amounts to nothing more than changing the units of length. Also note that from this scaling procedure, the absolute value of the largest element of \( u_k \) is always equal to 1. Therefore, for large \( m \),

\[
\begin{align*}
    u_m &= \frac{\lambda_n^m c_n x_n}{S_2 S_3 \cdots S_m} + \text{(relatively insignificant term)}.
\end{align*}
\]

Therefore, the entry of \( Au_m \) which has the largest absolute value is essentially equal to the entry having largest absolute value of

\[
A \left( \frac{\lambda_n^m c_n x_n}{S_2 S_3 \cdots S_m} \right) = \frac{\lambda_n^{m+1} c_n x_n}{S_2 S_3 \cdots S_m} \approx \lambda_n u_m
\]

and so for large \( m \), it must be the case that \( \lambda_n \approx S_m^{m+1} \). This suggests the following procedure.

**Finding the largest eigenvalue with its eigenvector.**

1. Start with a vector, \( u_1 \) which you hope has a component in the direction of \( x_n \). The vector, \((1, \cdots, 1)^T\) is usually a pretty good choice.

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.

**Example 32.1.1** Find the largest eigenvalue of \( A = \begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \).

The power method will now be applied to find the largest eigenvalue for the above matrix. Letting \( u_1 = (1, \cdots, 1)^T \), we will consider \( Au_1 \) and scale it.

\[
\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ -4 \\ 6 \end{pmatrix}.
\]

Scaling this vector by dividing by the largest entry gives

\[
\frac{1}{6} \begin{pmatrix} 2 \\ -4 \\ 6 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ 1 \end{pmatrix} = u_2
\]
Now let's do it again.

\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
\frac{1}{3} \\
\frac{-2}{3} \\
1
\end{pmatrix}
= 
\begin{pmatrix}
\frac{22}{1} \\
-8 \\
-6
\end{pmatrix}
\]

Then

\[
u_3 = \frac{1}{22}
\begin{pmatrix}
22 \\
-8 \\
-6
\end{pmatrix}
= 
\begin{pmatrix}
\frac{1}{11} \\
\frac{-4}{11} \\
\frac{-2}{11}
\end{pmatrix}
= 
\begin{pmatrix}
0.09090909 \\
-0.36363636 \\
-0.27272727
\end{pmatrix}
\]

Continue doing this

\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
1 \\
-0.36363636 \\
-0.27272727
\end{pmatrix}
= 
\begin{pmatrix}
7.090909 \\
-4.363636 \\
1.6363637
\end{pmatrix}
\]

Then

\[
u_4 = 
\begin{pmatrix}
1.012987 \\
-0.62337663 \\
0.23376624
\end{pmatrix}
\]

So far the scaling factors are changing fairly noticeably so continue.

\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
1.012987 \\
-0.62337663 \\
0.23376624
\end{pmatrix}
= 
\begin{pmatrix}
16.363636 \\
-7.4805195 \\
-1.4025975
\end{pmatrix}
\]

\[
u_5 = 
\begin{pmatrix}
.99778268 \\
-0.45612924 \\
-8.5524238 \times 10^{-2}
\end{pmatrix}
\]

\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
.99778268 \\
-0.45612924 \\
-8.5524238 \times 10^{-2}
\end{pmatrix}
= 
\begin{pmatrix}
10.433956 \\
-5.4735507 \\
.51314531
\end{pmatrix}
\]

\[
u_6 = \frac{1}{10}
\begin{pmatrix}
10.433956 \\
-5.4735507 \\
.51314531
\end{pmatrix}
= 
\begin{pmatrix}
1.0433956 \\
-.54735507 \\
.051314531 \times 10^{-2}
\end{pmatrix}
\]

\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
1.0433956 \\
-.54735507 \\
.051314531 \times 10^{-2}
\end{pmatrix}
= 
\begin{pmatrix}
13.444409 \\
-6.5682608 \\
-0.30788721
\end{pmatrix}
\]

\[
u_7 = 
\begin{pmatrix}
1.0341853 \\
-.50525083 \\
-2.3683632 \times 10^{-2}
\end{pmatrix}
\]

\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
1.0341853 \\
-.50525083 \\
-2.3683632 \times 10^{-2}
\end{pmatrix}
= 
\begin{pmatrix}
11.983918 \\
-6.06301 \\
.14210182
\end{pmatrix}
\]

\[
u_8 = 
\begin{pmatrix}
.99865983 \\
-.50525083 \\
1.1841818 \times 10^{-2}
\end{pmatrix}
\]
At this point, you could stop because the scaling factors are not changing by much. They went from 12.197 to 12.0904. It looks like the eigenvalue is something like 12 which is in fact the case. The eigenvector is approximately $u_{10}$. The true eigenvector for $\lambda = 12$ is
\[
\begin{pmatrix}
1 \\
-0.5 \\
0
\end{pmatrix}
\]
and so you see this is pretty close. If you didn’t know this, observe
\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
1.0075413 \\
-0.50525084 \\
2.960463 \times 10^{-3}
\end{pmatrix}
= 
\begin{pmatrix}
12.143783 \\
-6.0630104 \\
-1.7762529 \times 10^{-2}
\end{pmatrix}
\]
and
\[
12.090495 \begin{pmatrix}
1.0075413 \\
-0.50525084 \\
2.960463 \times 10^{-3}
\end{pmatrix}
= 
\begin{pmatrix}
12.181673 \\
-6.1087328 \\
3.5793463 \times 10^{-2}
\end{pmatrix}.
\]

### 32.2 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. 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 $\lambda$. 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, we 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. When using this method it is nice to choose $\alpha$ fairly close to an eigenvalue. Otherwise, the method will converge slowly. In order to get some idea where to start, you could use Gerschgorin’s theorem but this theorem will only give a rough idea where to look. There isn’t a really good way to know how to choose $\alpha$ for general cases. As we mentioned earlier, the eigenvalue problem is very difficult to solve in general.

**Lemma 32.2.1** 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 $\frac{1}{\lambda_k - \alpha}$.
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
\[
\frac{1}{\lambda_k - \alpha} x_k = (A - \alpha I)^{-1} x_k.
\]
This proves the lemma.

In explaining why the method works, we will assume \( A \) is nondefective. This is not necessary! One can use Gelfand’s theorem on the spectral radius which is presented in the appendix titled Heroic Linear Algebra and invariance of \((A - \alpha I)^{-1}\) on generalized eigenspaces to prove a more general result than what we will present here. It suffices to assume that the eigenspace for \( \lambda_k \) has dimension equal to the multiplicity of the eigenvalue \( \lambda_k \). The point is, this method is much better than it might seem from the explanation we are about to give. Pick \( u_1 \), an initial vector and let \( A x_k = \lambda_k x_k \), where \( \{x_1, \ldots, x_n\} \) is a basis of eigenvectors which exists from the assumption that \( A \) is nondefective. Assume \( \alpha \) is closer to \( \lambda_n \) than to any other eigenvalue. Since \( A \) is nondefective, there exist constants, \( a_k \) such that
\[
u_1 = \sum_{k=1}^{n} a_k x_k.
\]
Possibly \( \lambda_n \) is a repeated eigenvalue. Then combining the terms in the sum which involve eigenvectors for \( \lambda_n \), a simpler description of \( u_1 \) is
\[
u_1 = \sum_{j=1}^{m} a_j x_j + y
\]
where \( y \) is an eigenvector for \( \lambda_n \) which is assumed not equal to \( 0 \). (If you are unlucky in your choice for \( u_1 \), this might not happen and things won’t work.) Now the iteration procedure is defined as
\[
u_{k+1} = (A - \alpha I)^{-1} \nu_k
\]
where \( S_k \) is the element of \((A - \alpha I)^{-1} \nu_k\) which has largest absolute value. From Lemma 32.2.1,
\[
u_{k+1} = \frac{\sum_{j=1}^{m} a_j \left( \frac{1}{\lambda_j - \alpha} \right)^k x_j + \left( \frac{1}{\lambda_n - \alpha} \right)^k y}{S_2 \cdots S_k}
\]
Now it is being assumed that \( \lambda_n \) is the eigenvalue which is closest to \( \alpha \) and so for large \( k \), the term,
\[
\sum_{j=1}^{m} a_j \left( \frac{\lambda_n - \alpha}{\lambda_j - \alpha} \right)^k x_j \equiv E_k
\]
is very small while for every \( k \geq 1 \), \( u_k \) is a moderate sized vector because every entry has absolute value less than or equal to 1. Thus
\[
u_{k+1} = \frac{\left( \frac{1}{\lambda_n - \alpha} \right)^k}{S_2 \cdots S_k} (E_k + y) \equiv C_k (E_k + y)
\]
where $E_k \to 0$, $y$ is some eigenvector for $\lambda_n$, and $C_k$ is of moderate size, remaining bounded as $k \to \infty$. Therefore, for large $k$,

$$u_{k+1} - C_ky = C_kE_k \approx 0$$

and multiplying by $(A - \alpha I)^{-1}$ yields

$$(A - \alpha I)^{-1}u_{k+1} - (A - \alpha I)^{-1}C_ky = (A - \alpha I)^{-1}u_{k+1} - C_k\left(\frac{1}{\lambda_n - \alpha}\right)y$$

$$\approx (A - \alpha I)^{-1}u_{k+1} - \left(\frac{1}{\lambda_n - \alpha}\right)u_{k+1} \approx 0.$$ 

Therefore, for large $k$, $u_k$ is approximately equal to an eigenvector of $(A - \alpha I)^{-1}$. Therefore,

$$(A - \alpha I)^{-1}u_k \approx \frac{1}{\lambda_n - \alpha}u_k$$

and so you could take the dot product of both sides with $u_k$ and approximate $\lambda_n$ by solving the following for $\lambda_n$.

$$\frac{(A - \alpha I)^{-1}u_k \cdot u_k}{|u_k|^2} = \frac{1}{\lambda_n - \alpha}$$

How else can you find the eigenvalue from this? Suppose $u_k = (w_1, \cdots, w_n)^T$ and from the construction $|w_i| \leq 1$ and $w_k = 1$ for some $k$. Then

$$S_ku_{k+1} = (A - \alpha I)^{-1}u_k \approx (A - \alpha I)^{-1}(C_{k-1}y) = \frac{1}{\lambda_n - \alpha}(C_{k-1}y) \approx \frac{1}{\lambda_n - \alpha}u_k.$$ 

Hence the entry of $(A - \alpha I)^{-1}u_k$ which has largest absolute value is approximately $\frac{1}{\lambda_n - \alpha}$ and so it is likely that you can estimate $\lambda_n$ using the formula

$$S_k = \frac{1}{\lambda_n - \alpha}.$$ 

Of course this would fail if $(A - \alpha I)^{-1}u_k$ had more than one entry having equal absolute value.

**Here is how you use the shifted inverse power method to find the eigenvalue and eigenvector closest to $\alpha$**.

1. Find $(A - \alpha I)^{-1}$.

2. Pick $u_1$. It is important that $u_1 = \sum_{j=1}^m a_jx_j + y$ where $y$ is an eigenvector which goes with the eigenvalue closest to $\alpha$ and the sum is in an “invariant subspace corresponding to the other eigenvalues”. Of course you have no way of knowing whether this is so but it typically is so. If things don’t work out, just start with a different $u_1$. You were unlucky in your choice.

3. If $u_k$ has been obtained,

$$u_{k+1} = \frac{(A - \alpha I)^{-1}u_k}{S_k}$$

where $S_k$ is the element of $u_k$ 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

\[
S_k = \frac{1}{\lambda - \alpha}
\]

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.

**Example 32.2.2** 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.

In this case the eigenvalues are \(-6, 0, \) and \(12\) so the correct answer is \(-6\) for the eigenvalue. Then from the above procedure, we will start with an initial vector,

\[
u_1 \equiv \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.
\]

Then we must solve the following equation.

\[
\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} + 7 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}
\]

Simplifying the matrix on the left, we must solve

\[
\begin{pmatrix} 12 & -14 & 11 \\ -4 & 11 & -4 \\ 3 & 6 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}
\]

and then divide by the entry which has largest absolute value to obtain

\[
u_2 = \begin{pmatrix} 1.0 \\ 0.184 \\ -0.76 \end{pmatrix}
\]

Now solve

\[
\begin{pmatrix} 12 & -14 & 11 \\ -4 & 11 & -4 \\ 3 & 6 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1.0 \\ 0.184 \\ -0.76 \end{pmatrix}
\]

and divide by the largest entry, 1.0515 to get

\[
u_3 = \begin{pmatrix} 1.0 \\ 0.266 \\ -0.97061 \end{pmatrix}
\]

Solve

\[
\begin{pmatrix} 12 & -14 & 11 \\ -4 & 11 & -4 \\ 3 & 6 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1.0 \\ 0.266 \\ -0.97061 \end{pmatrix}
\]
and divide by the largest entry, 1.01 to get

\[ \mathbf{u}_4 = \left( \begin{array}{c} 1.0 \\ 3.8454 \times 10^{-3} \\ -0.99604 \end{array} \right). \]

These scaling factors are pretty close after these few iterations. Therefore, the predicted eigenvalue is obtained by solving the following for \( \lambda \).

\[ \frac{1}{\lambda + 7} = 1.01 \]

which gives \( \lambda = -6.01 \). You see this is pretty close. In this case the eigenvalue closest to \(-7\) was \(-6\).

**Example 32.2.3** Consider the symmetric matrix, 

\[ A = \left( \begin{array}{ccc} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{array} \right). \]

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

\[ p(\lambda) = \det(\lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - A) = \lambda^3 - 4\lambda^2 - 24\lambda - 17 = 0 \]

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. Of course you could zoom in and find it very accurately without much trouble but what about the eigenvector which goes with it? If you try to solve

\[ \left( \begin{array}{ccc} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{array} \right) \left( \begin{array}{c} x \\ y \\ z \end{array} \right) = \left( \begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right) \]

there will be only the zero solution because the matrix on the left will be invertible and the same will be true if you replace \(-.8\) with a better approximation like \(-.86\) or \(-.855\). This is because all these are only approximations to the eigenvalue and so the matrix in the above is nonsingular for all of these. Therefore, you will only get the zero solution and

**Eigenvectors are never equal to zero!**

However, there exists such an eigenvector and you can find it using the shifted inverse power method. Pick \( \alpha = -.855 \). Then you solve

\[ \left( \begin{array}{ccc} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{array} \right) + .855 \left( \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right) \left( \begin{array}{c} x \\ y \\ z \end{array} \right) = \left( \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \right) \]

or in other words,

\[ \left( \begin{array}{ccc} 1.855 & 2.0 & 3.0 \\ 2.0 & 1.855 & 4.0 \\ 3.0 & 4.0 & 2.855 \end{array} \right) \left( \begin{array}{c} x \\ y \\ z \end{array} \right) = \left( \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \right). \]
Divide by the largest entry, $-67.944$, to obtain

$$u_2 = \begin{pmatrix} 1.0 \\ -0.58921 \\ -0.23044 \end{pmatrix}. $$

Now solve

$$ \begin{pmatrix} 1.855 & 2.0 & 3.0 \\ 2.0 & 1.855 & 4.0 \\ 3.0 & 4.0 & 2.855 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1.0 \\ -0.58921 \\ -0.23044 \end{pmatrix}. $$

The solution is:

$$ u_3 = \begin{pmatrix} -514.01 \\ 302.12 \\ 116.75 \end{pmatrix} $$

and divide by the largest entry, $-514.01$, to obtain

$$ u_3 = \begin{pmatrix} 1.0 \\ -0.58777 \\ -0.22714 \end{pmatrix}. $$

Clearly the $u_k$ are not changing much. This suggests an approximate eigenvector for this eigenvalue which is close to $-0.855$ is the above $u_3$. And an eigenvalue is obtained by solving

$$ \frac{1}{\lambda + 0.855} = -514.01 $$

$\lambda = -0.8569$. Let's check this.

$$ \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix} \begin{pmatrix} -0.58777 \\ -0.22714 \end{pmatrix} = \begin{pmatrix} -0.8569 \\ -0.5037 \\ 0.1946 \end{pmatrix}. $$

Thus the vector of (32.2) is very close to the desired eigenvector, just as $-0.8569$ is very close to the desired eigenvalue. For practical purposes, we have found both the eigenvector and the eigenvalue.

**Example 32.2.4** Find the eigenvalues and eigenvectors of the matrix, $A = \begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix}$.

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 $-0.4$, 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.

Let's first try to find the eigenvector and a better 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 for the first iteration, letting $u_1 = (1, 1, 1)^T$,

\[
\begin{pmatrix}
-25.357143 & -33.928571 & 50.0 \\
12.5 & 17.5 & -25.0 \\
23.214286 & 30.357143 & -45.0
\end{pmatrix}
\begin{pmatrix}
1 \\
1 \\
1
\end{pmatrix}
= \begin{pmatrix}
-9.285714 \\
5.0 \\
8.571429
\end{pmatrix}
\]

To get $u_2$, we must divide by $-9.285714$. Thus

\[
u_2 = \begin{pmatrix}
1.0 \\
-0.53846156 \\
-0.923077
\end{pmatrix}.
\]

Do another iteration.

\[
\begin{pmatrix}
-25.357143 & -33.928571 & 50.0 \\
12.5 & 17.5 & -25.0 \\
23.214286 & 30.357143 & -45.0
\end{pmatrix}
\begin{pmatrix}
1.0 \\
-0.53846156 \\
-0.923077
\end{pmatrix}
= \begin{pmatrix}
-53.241762 \\
26.153848 \\
48.406596
\end{pmatrix}
\]

Then to get $u_3$ you divide by $-53.241762$. Thus

\[
u_3 = \begin{pmatrix}
1.0 \\
-0.49122807 \\
-0.90918471
\end{pmatrix}.
\]

Now iterate again because the scaling factors are still changing quite a bit.

\[
\begin{pmatrix}
-25.357143 & -33.928571 & 50.0 \\
12.5 & 17.5 & -25.0 \\
23.214286 & 30.357143 & -45.0
\end{pmatrix}
\begin{pmatrix}
1.0 \\
-0.49122807 \\
-0.90918471
\end{pmatrix}
= \begin{pmatrix}
-54.149712 \\
26.633127 \\
49.215317
\end{pmatrix}.
\]

This time the scaling factor didn’t change too much. It is $-54.149712$. Thus

\[
u_4 = \begin{pmatrix}
1.0 \\
-0.49184245 \\
-0.90887495
\end{pmatrix}.
\]

Lets do one more iteration.

\[
\begin{pmatrix}
-25.357143 & -33.928571 & 50.0 \\
12.5 & 17.5 & -25.0 \\
23.214286 & 30.357143 & -45.0
\end{pmatrix}
\begin{pmatrix}
1.0 \\
-0.49184245 \\
-0.90887495
\end{pmatrix}
= \begin{pmatrix}
-54.113379 \\
26.614631 \\
49.182727
\end{pmatrix}.
\]

You see at this point the scaling factors have definitely settled down and so it seems our eigenvalue would be obtained by solving

\[
\frac{1}{\lambda - (-1.2)} = -54.113379
\]

and this yields $\lambda = -1.2184797$ as an approximation to the eigenvalue and the eigenvector would be obtained by dividing by $-54.113379$ which gives

\[
u_5 = \begin{pmatrix}
1.0000002 \\
-0.49183097 \\
-0.90888309
\end{pmatrix}.
\]

How well does it work?

\[
\begin{pmatrix}
2 & 1 & 3 \\
2 & 1 & 1 \\
3 & 2 & 1
\end{pmatrix}
\begin{pmatrix}
1.0000002 \\
-0.49183097 \\
-0.90888309
\end{pmatrix}
= \begin{pmatrix}
-1.2184798 \\
0.59928634 \\
1.1074556
\end{pmatrix}.
\]
while

\[
-1.2184797 \begin{pmatrix} 1.0000002 \\ -0.49183097 \\ -0.90888309 \end{pmatrix} = \begin{pmatrix} -1.2184799 \\ 0.59928605 \\ 1.1074556 \end{pmatrix}.
\]

For practical purposes, this has found the eigenvalue near \(-1.2\) as well as an eigenvector associated with it.

Next we shall find the eigenvector and a more precise value for the eigenvalue near \(-.4\). In this case,

\[
(A - \alpha I)^{-1} = \begin{pmatrix} 8.0645161 \times 10^{-2} & -9.2741935 & 6.4516129 \\ -0.40322581 & 11.370968 & -7.2580645 \\ 0.40322581 & 3.6290323 & -2.7419355 \end{pmatrix}.
\]

As before, we have no idea what the eigenvector is so we will again try \((1,1,1)^T\). Then to find \(u_2\),

\[
\begin{pmatrix} 8.0645161 \times 10^{-2} & -9.2741935 & 6.4516129 \\ -0.40322581 & 11.370968 & -7.2580645 \\ 0.40322581 & 3.6290323 & -2.7419355 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -2.7419354 \\ 3.7096777 \\ 1.2903226 \end{pmatrix}.
\]

The scaling factor is 3.709677. Thus

\[
u_2 = \begin{pmatrix} 0.73913036 \\ 1.0 \\ 0.34782607 \end{pmatrix}.
\]

Now lets do another iteration.

\[
\begin{pmatrix} 8.0645161 \times 10^{-2} & -9.2741935 & 6.4516129 \\ -0.40322581 & 11.370968 & -7.2580645 \\ 0.40322581 & 3.6290323 & -2.7419355 \end{pmatrix} \begin{pmatrix} 0.73913036 \\ 1.0 \\ 0.34782607 \end{pmatrix} = \begin{pmatrix} -7.0897616 \\ 9.1444604 \\ 2.3772792 \end{pmatrix}.
\]

The scaling factor is 9.144460. Thus

\[
u_3 = \begin{pmatrix} -0.77530672 \\ 1.0 \\ 0.25996933 \end{pmatrix}.
\]

Lets do another iteration. The scaling factors are still changing quite a bit.

\[
\begin{pmatrix} 8.0645161 \times 10^{-2} & -9.2741935 & 6.4516129 \\ -0.40322581 & 11.370968 & -7.2580645 \\ 0.40322581 & 3.6290323 & -2.7419355 \end{pmatrix} \begin{pmatrix} -0.77530672 \\ 1.0 \\ 0.25996933 \end{pmatrix} = \begin{pmatrix} -7.6594968 \\ 9.7967175 \\ 2.6035895 \end{pmatrix}.
\]

The scaling factor is now 9.7967175. Therefore,

\[
u_4 = \begin{pmatrix} -0.78184318 \\ 1.0 \\ 0.26576141 \end{pmatrix}.
\]

Lets do another iteration.

\[
\begin{pmatrix} 8.0645161 \times 10^{-2} & -9.2741935 & 6.4516129 \\ -0.40322581 & 11.370968 & -7.2580645 \\ 0.40322581 & 3.6290323 & -2.7419355 \end{pmatrix} \begin{pmatrix} -0.78184318 \\ 1.0 \\ 0.26576141 \end{pmatrix} = \begin{pmatrix} -7.6226556 \\ 9.7573139 \\ 2.5850723 \end{pmatrix}.
\]
Now the scaling factor is 9.7573139 and so

\[ u_5 = \begin{pmatrix} -.7812248 \\ 1.0 \\ .26493688 \end{pmatrix}. \]

We notice the scaling factors are not changing by much so the approximate eigenvalue is

\[ \frac{1}{\lambda + .4} = 9.7573139 \]

which shows \( \lambda = -.29751278 \) is an approximation to the eigenvalue near \( .4 \). How well does it work?

\[ \begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} -.7812248 \\ 1.0 \\ .26493688 \end{pmatrix} = \begin{pmatrix} .23236104 \\ -.29751272 \\ -.07873752 \end{pmatrix}. \]

\[ \begin{pmatrix} -.7812248 \\ 1.0 \\ .26493688 \end{pmatrix} \begin{pmatrix} .23242436 \\ -.29751278 \\ -7.8822108 \times 10^{-2} \end{pmatrix}. \]

It works pretty well. For practical purposes, the eigenvalue and eigenvector have now been found. If you want better accuracy, you could just continue iterating.

Next we will find the eigenvalue and eigenvector for the eigenvalue near 5.5. In this case,

\[ (A - \alpha I)^{-1} = \begin{pmatrix} 29.2 & 16.8 & 23.2 \\ 19.2 & 10.8 & 15.2 \\ 28.0 & 16.0 & 22.0 \end{pmatrix}. \]

As before, we have no idea what the eigenvector is but we don’t want to give the impression that you always need to start with the vector \((1,1,1)^T\). Therefore, we shall let \( u_1 = (1,2,3)^T \). What follows is the iteration without all the comments between steps.

\[ S_1 = 86.4. \]

\[ u_2 = \begin{pmatrix} 1.5324074 \\ 1.0 \\ 1.4583333 \end{pmatrix}. \]

\[ \begin{pmatrix} 29.2 & 16.8 & 23.2 \\ 19.2 & 10.8 & 15.2 \\ 28.0 & 16.0 & 22.0 \end{pmatrix} \begin{pmatrix} 1.5324074 \\ 1.0 \\ 1.4583333 \end{pmatrix} = \begin{pmatrix} 95.379629 \\ 62.388888 \\
90.99074 \end{pmatrix}. \]

\[ S_2 = 95.379629. \]

\[ u_3 = \begin{pmatrix} 1.0 \\ .65411125 \\ .95398505 \end{pmatrix}. \]

\[ \begin{pmatrix} 29.2 & 16.8 & 23.2 \\ 19.2 & 10.8 & 15.2 \\ 28.0 & 16.0 & 22.0 \end{pmatrix} \begin{pmatrix} 1.0 \\ .65411125 \\ .95398505 \end{pmatrix} = \begin{pmatrix} 62.321522 \\ 40.764974 \\
59.453451 \end{pmatrix}. \]
32.2. THE SHIFTED INVERSE POWER METHOD

\[ S_3 = 62.321522. \]

\[ u_4 = \begin{pmatrix} 1.0 \\ .65410748 \\ .95397945 \end{pmatrix} \]

\[
\begin{pmatrix}
29.2 & 16.8 & 23.2 \\
19.2 & 10.8 & 15.2 \\
28.0 & 16.0 & 22.0
\end{pmatrix}
\begin{pmatrix} 1.0 \\ .65410748 \\ .95397945 \end{pmatrix} = \begin{pmatrix} 62.321329 \\ 40.764848 \\ 59.453268 \end{pmatrix}
\]

\[ S_4 = 62.321329. \] Looks like it is time to stop because this scaling factor is not changing much from \( S_3. \)

\[ u_5 = \begin{pmatrix} 1.0 \\ .65410749 \\ .95397946 \end{pmatrix}. \]

Then the approximation of the eigenvalue is gotten by solving

\[ 62.321329 = \frac{1}{\lambda - 5.5} \]

which gives \( \lambda = 5.5160459. \) Lets see how well it works.

\[
\begin{pmatrix}
2 & 1 & 3 \\
2 & 1 & 1 \\
3 & 2 & 1
\end{pmatrix}
\begin{pmatrix} 1.0 \\ .65410749 \\ .95397946 \end{pmatrix} = \begin{pmatrix} 5.5160459 \\ 3.608087 \\ 5.2621944 \end{pmatrix}
\]

\[
5.5160459
\begin{pmatrix} 1.0 \\ .65410749 \\ .95397946 \end{pmatrix} = \begin{pmatrix} 5.5160459 \\ 3.6080869 \\ 5.2621945 \end{pmatrix}. \]

32.2.1 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 32.2.5** 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 \( 1 - i. \) We will apply the process to \( \alpha = i \) so we will find the eigenvalue closest to \( i. \)

\[
(A - \alpha I)^{-1} = \begin{pmatrix}
-.02 - .14i & 1.24 + .68i & -.84 + .12i \\
-.14 + .02i & .68 - .24i & .12 + .84i \\
.02 + .14i & -.24 - .68i & .84 + .88i
\end{pmatrix}
\]
Then let $u_1 = (1, 1, 1)^T$ for lack of any insight into anything better.

$$
\begin{pmatrix}
-0.2 - .14i & 1.24 + .68i & - .84 + .12i \\
-1.14 + .02i & .68 - .24i & .12 + .84i \\
.02 + .14i & -.24 - .68i & .84 + .88i \\
\end{pmatrix}
\begin{pmatrix}
1 \\
1 \\
1 \\
\end{pmatrix}
= 
\begin{pmatrix}
.38 + .66i \\
.66 + .62i \\
.62 + .34i \\
\end{pmatrix}
$$

$S_2 = .66 + .62i$.

$$
u_2 = 
\begin{pmatrix}
.804 878 05 + .243 902 44i \\
1.0 \\
.756 097 56 - .195 121 95i
\end{pmatrix}
= 
\begin{pmatrix}
-0.2 - .14i & 1.24 + .68i & - .84 + .12i \\
-1.14 + .02i & .68 - .24i & .12 + .84i \\
.02 + .14i & -.24 - .68i & .84 + .88i \\
\end{pmatrix}
\begin{pmatrix}
.804 878 05 + .243 902 44i \\
1.0 \\
.756 097 56 - .195 121 95i
\end{pmatrix}
$$

$S_3 = .646 341 46 + .817 073 17i$. After more iterations, of this sort, you find $S_9 = 1.002 748 5 + 2.137 621 7 \times 10^{-4}i$ and

$$
u_9 = 
\begin{pmatrix}
1.0 \\
.501 514 17 - .499 807 33i \\
1.562 088 1 \times 10^{-3} - .499 778 55i
\end{pmatrix}
$$

Then

$$
\begin{pmatrix}
-0.2 - .14i & 1.24 + .68i & - .84 + .12i \\
-1.14 + .02i & .68 - .24i & .12 + .84i \\
.02 + .14i & -.24 - .68i & .84 + .88i \\
\end{pmatrix}
\begin{pmatrix}
1.0 \\
.501 514 17 - .499 807 33i \\
1.562 088 1 \times 10^{-3} - .499 778 55i
\end{pmatrix}
$$

$S_{10} = 1.000 407 8 + 1.269 979 \times 10^{-3}i$.

$$
u_{10} = 
\begin{pmatrix}
1.0 \\
.500 239 18 - .499 325 33i \\
2.506 749 2 \times 10^{-4} - .499 311 92i
\end{pmatrix}
$$

The scaling factors are not changing much at this point

$$1.000 407 8 + 1.269 979 \times 10^{-3}i = \frac{1}{\lambda - i}$$

The approximate eigenvalue is then $\lambda = .999 590 76 + .998 731 06i$. This is pretty close to $1 + i$. How well does the eigenvector work?

$$
\begin{pmatrix}
5 & -8 & 6 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
\end{pmatrix}
\begin{pmatrix}
1.0 \\
.500 239 18 - .499 325 33i \\
2.506 749 2 \times 10^{-4} - .499 311 92i
\end{pmatrix}
= 
\begin{pmatrix}
.999 590 61 + .998 731 12i \\
500 239 18 - .499 325 33i \\
1.0
\end{pmatrix}
$$
32.3. The Rayleigh Quotient

The Rayleigh Quotient

\[
\begin{pmatrix}
0.9959076 + 0.99873106i \\
0.99872618 + 4.8342039 \times 10^{-4}i
\end{pmatrix}
\]

\[
\begin{pmatrix}
1.0 \\
2.5067492 \times 10^{-4} - 0.49931192i
\end{pmatrix}
\]

\[
= \begin{pmatrix}
0.99959076 + 0.99873106i \\
0.99872618 + 4.8342039 \times 10^{-4}i
\end{pmatrix}
\]

\[
= \begin{pmatrix}
-0.99959076 + 0.99873106i \\
-0.99872618 + 4.8342039 \times 10^{-4}i
\end{pmatrix}
\]

It took more iterations than before because \( \alpha \) was not very close to \( 1 + i \). 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?

32.3 The Rayleigh Quotient

There are many specialized results concerning the eigenvalues and eigenvectors for Hermitian matrices. 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 \( x \in \mathbb{F}^n \),

\[
|x|^2 = x^*x = \sum_{j=1}^{n} |x_j|^2.
\]

The following corollary gives the theoretical foundation for the spectral theory of Hermitian matrices it was proved in the section on Schur’s theorem.

Corollary 32.3.1 If \( A \) is Hermitian, then all the eigenvalues of \( A \) are real and there exists an orthonormal basis of eigenvectors.

Thus for \( \{x_k\}_{k=1}^{n} \) this orthonormal basis,

\[
x^*_k x_j = \delta_{ij} \equiv \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}
\]

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 $A x_k = \lambda_k x_k$ 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,

$$|x|^2 = \sum_{i=1}^n \overline{a_i} x_i^* \sum_{j=1}^n a_j x_j$$

$$= \sum_{ij} \overline{a_i} a_j x_i^* x_j = \sum_{ij} \overline{a_i} a_j \delta_{ij} = \sum_{i=1}^n |a_i|^2.$$ 

Therefore,

$$\frac{x^* A x}{|x|^2} = \frac{\left(\sum_{i=1}^n \overline{a_i} x_i^* \right) \left(\sum_{j=1}^n a_j \lambda_j x_j\right)}{\sum_{i=1}^n |a_i|^2}$$

$$= \sum_{ij} \overline{a_i} a_j \lambda_j x_i^* x_j = \sum_{ij} \overline{a_i} a_j \lambda_j \delta_{ij}$$

$$= \sum_{i=1}^n |a_i|^2 \lambda_i \sum_{i=1}^n \frac{|a_i|^2}{|a_i|^2} \in [\lambda_1, \lambda_n].$$ 

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 32.3.2** Let $x \neq 0$ and form the Rayleigh quotient,

$$\frac{x^* A x}{|x|^2} \equiv q.$$ 

Then there exists an eigenvalue of $A$, denoted here by $\lambda_q$ such that

$$|\lambda_q - q| \leq \frac{|A x - q x|}{|x|}. \quad (32.3)$$

**Proof:** Let $x = \sum_{k=1}^n a_k x_k$ where $\{x_k\}_{k=1}^n$ is the orthonormal basis of eigenvectors.

$$|A x - q x|^2 = (A x - q x)^* (A x - q x)$$

$$= \left(\sum_{k=1}^n a_k \lambda_k x_k - q a_k x_k\right)^* \left(\sum_{k=1}^n a_k \lambda_k x_k - q a_k x_k\right)$$

$$= \left(\sum_{j=1}^n (\lambda_j - q) \overline{a_j} x_j^* \right) \left(\sum_{k=1}^n (\lambda_k - q) a_k x_k\right)$$

$$= \sum_{j,k} (\lambda_j - q) \overline{a_j} (\lambda_k - q) a_k x_j^* x_k$$

$$= \sum_{k=1}^n |a_k|^2 (\lambda_k - q)^2.$$
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 (32.3).

**Example 32.3.3** 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

\[
\left( \begin{array}{ccc} 1 & 1 & 1 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{array} \right) \left( \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \right) = \frac{19}{3}
\]

According to the above theorem, there is some eigenvalue of this matrix, \( \lambda_q \) such that

\[
\left| \lambda_q - \frac{19}{3} \right| \leq \frac{\left| \left( \begin{array}{ccc} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{array} \right) \left( \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \right) - \frac{19}{3} \left( \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \right) \right|}{\sqrt{3}} = \frac{1}{\sqrt{3}} \left( \begin{array}{c} -\frac{4}{9} \\ -\frac{5}{9} \end{array} \right)
\]

\[
= \frac{\sqrt{\frac{1}{9} + \left( \frac{4}{9} \right)^2 + \left( \frac{5}{9} \right)^2}}{\sqrt{3}} = 1.2472
\]

Could you find this eigenvalue and associated eigenvector? Of course you could. This is what the inverse shifted power method is all about.

Solve

\[
\left( \begin{array}{ccc} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{array} \right) - \frac{19}{3} \left( \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right) \left( \begin{array}{c} x \\ y \\ z \end{array} \right) = \left( \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \right)
\]

In other words solve

\[
\left( \begin{array}{ccc} -\frac{16}{9} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{array} \right) \left( \begin{array}{c} x \\ y \\ z \end{array} \right) = \left( \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \right)
\]

and divide by the entry which is largest, 3.8707, to get

\[
u_2 = \begin{pmatrix} .69925 \\ .49389 \\ 1.0 \end{pmatrix}
\]

Now solve

\[
\left( \begin{array}{ccc} -\frac{16}{9} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{array} \right) \left( \begin{array}{c} x \\ y \\ z \end{array} \right) = \begin{pmatrix} .69925 \\ .49389 \\ 1.0 \end{pmatrix}
\]
and divide by the entry with largest absolute value, 2.9979 to get
\[ u_3 = \begin{pmatrix} 0.71473 \\ 0.52263 \\ 1.0 \end{pmatrix} \]

Now solve
\[
\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0.71473 \\ 0.52263 \\ 1.0 \end{pmatrix}
\]

and divide by the entry with largest absolute value, 3.0454, to get
\[ u_4 = \begin{pmatrix} 0.7137 \\ 0.52056 \\ 1.0 \end{pmatrix} \]

Solve
\[
\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0.7137 \\ 0.52056 \\ 1.0 \end{pmatrix}
\]

and divide by the largest entry, 3.0421 to get
\[ u_5 = \begin{pmatrix} 0.71378 \\ 0.52073 \\ 1.0 \end{pmatrix} \]

You can see these scaling factors are not changing much. The predicted eigenvalue is obtained by solving
\[ \frac{1}{\lambda - \frac{19}{3}} = 3.0421 \]

to obtain \( \lambda = 6.6621 \). How close is this?
\[
\begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix} \begin{pmatrix} 0.71378 \\ 0.52073 \\ 1.0 \end{pmatrix} = \begin{pmatrix} 4.7552 \\ 3.4691 \\ 6.6621 \end{pmatrix}
\]

while
\[
6.6621 \begin{pmatrix} 0.71378 \\ 0.52073 \\ 1.0 \end{pmatrix} = \begin{pmatrix} 4.7553 \\ 3.4692 \\ 6.6621 \end{pmatrix}.
\]

You see that for practical purposes, this has found the eigenvalue and an eigenvector.

### 32.4 Exercises

1. In Example 32.3.3 an eigenvalue was found correct to several decimal places along with an eigenvector. Find the other eigenvalues along with their eigenvectors.

2. Using the power method, find the eigenvalue correct to one decimal place having largest absolute value for the matrix, \( A = \begin{pmatrix} 0 & -4 & -4 \\ 7 & 10 & 5 \\ -2 & 0 & 6 \end{pmatrix} \) along with an eigenvector associated with this eigenvalue.
3. Using the power method, find the eigenvalue correct to one decimal place having largest absolute value for the matrix, \( A = \begin{pmatrix} 15 & 6 & 1 \\ -5 & 2 & 1 \\ 1 & 2 & 7 \end{pmatrix} \) along with an eigenvector associated with this eigenvalue.

4. Using the power method, find the eigenvalue correct to one decimal place having largest absolute value for the matrix, \( A = \begin{pmatrix} 10 & 4 & 2 \\ -3 & 2 & -1 \\ 0 & 0 & 4 \end{pmatrix} \) along with an eigenvector associated with this eigenvalue.

5. Using the power method, find the eigenvalue correct to one decimal place having largest absolute value for the matrix, \( A = \begin{pmatrix} 15 & 14 & -3 \\ -13 & -18 & 9 \\ 5 & 10 & -1 \end{pmatrix} \) along with an eigenvector associated with this eigenvalue.

6. 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.

7. 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?

8. Find the eigenvalues and eigenvectors of the matrix, \( A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 0 & 3 \\ 1 & 3 & 2 \end{pmatrix} \) numerically. We don’t know the exact eigenvalues in this case. Check your answers by multiplying your numerically computed eigenvectors by the matrix.

9. Find the eigenvalues and eigenvectors of the matrix, \( A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 0 & 3 \\ 1 & 3 & 2 \end{pmatrix} \) numerically. We don’t know the exact eigenvalues in this case. Check your answers by multiplying your numerically computed eigenvectors by the matrix.

10. Consider the matrix, \( A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 0 \end{pmatrix} \) and the vector \((1,1,1)^T\). Estimate the distance between the Rayleigh quotient determined by this vector and some eigenvalue of \( A \).

11. Consider the matrix, \( A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 1 & 4 \\ 1 & 4 & 5 \end{pmatrix} \) and the vector \((1,1,1)^T\). Estimate the distance between the Rayleigh quotient determined by this vector and some eigenvalue of \( A \).
12. Consider the matrix, $A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 6 & 4 \\ 3 & 4 & -3 \end{pmatrix}$ and the vector $(1, 1, 1)^T$. Estimate the distance between the Rayleigh quotient determined by this vector and some eigenvalue of $A$.

13. Using Gershgorin’s theorem, find upper and lower bounds for the eigenvalues of $A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 6 & 4 \\ 3 & 4 & -3 \end{pmatrix}$. 
A.1 Worked Exercises

1. Here is an augmented matrix in which $*$ denotes an arbitrary number and ■ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

$$
\begin{pmatrix}
\text{■} & * & * & * & | & * \\
0 & \text{■} & * & 0 & | & * \\
0 & 0 & \text{■} & * & | & \text{■} \\
0 & 0 & 0 & 0 & | & * \\
\end{pmatrix}
$$

In this case the system is consistent and there is an infinite set of solutions. To see it is consistent, the bottom equation would yield a unique solution for $x_5$. Then letting $x_4 = t$, and substituting in to the other equations, beginning with the equation determined by the third row and then proceeding up to the next row followed by the first row, you get a solution for each value of $t$. There is a free variable which comes from the fourth column which is why you can say $x_4 = t$. Therefore, the solution is infinite.

2. Here is an augmented matrix in which $*$ denotes an arbitrary number and ■ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

$$
\begin{pmatrix}
\text{■} & * & * & | & * \\
0 & 0 & \text{■} & | & \text{■} \\
0 & 0 & * & | & 0 \\
\end{pmatrix}
$$

In this case there is no solution because you could use a row operation to place a 0 in the third row and third column position, like this:

$$
\begin{pmatrix}
\text{■} & * & * & | & * \\
0 & 0 & \text{■} & | & \text{■} \\
0 & 0 & 0 & | & \text{■} \\
\end{pmatrix}
$$

This would give a row of zeros equal to something nonzero.

3. Find $h$ such that

$$
\begin{pmatrix}
1 & h & | & 4 \\
3 & 7 & | & 7 \\
\end{pmatrix}
$$
is the augmented matrix of an inconsistent matrix.  

Doing a row operation by taking \(-3\) times the top row and adding to the bottom, this gives 

\[
\begin{pmatrix}
1 & h & | & 4 \\
0 & 7 - 3h & | & 7 - 12
\end{pmatrix}
\]

The system will be inconsistent if \(7 - 3h = 0\) or in other words, \(h = 7/3\).

4. Determine if the system is consistent.

\[
\begin{align*}
x + 2y + 3z - w &= 2 \\
x - y + 2z + w &= 1 \\
2x + 3y - z &= 1 \\
4x + 2y + z &= 5
\end{align*}
\]

The augmented matrix is

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

A reduced echelon form for this is

\[
\begin{pmatrix}
9 & 0 & 0 & 0 & | & 14 \\
0 & 9 & 0 & 0 & | & -6 \\
0 & 0 & 9 & 0 & | & 1 \\
0 & 0 & 0 & 9 & | & -13
\end{pmatrix}
\]

Therefore, there is a unique solution. In particular the system is consistent.

5. Find the point, \((x_1, y_1)\) which lies on both lines, \(5x + 3y = 1\) and \(4x - y = 3\).

You solve the system of equations whose augmented matrix is

\[
\begin{pmatrix}
5 & 3 & | & 1 \\
4 & -1 & | & 3
\end{pmatrix}
\]

A reduced echelon form is

\[
\begin{pmatrix}
17 & 0 & 10 \\
0 & 17 & -11
\end{pmatrix}
\]

and so the solution is \(x = 17/10\) and \(y = -11/17\).

6. Do the three lines, \(3x + 2y = 1\), \(2x - y = 1\), and \(4x + 3y = 3\) have a common point of intersection? If so, find the point and if not, tell why they don’t have such a common point of intersection.

This is asking for the solution to the three equations shown. The augmented matrix is

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

A reduced echelon form is

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

and this would require \(0x + 0y = 1\) which is impossible so there is no solution to this system of equations and hence no point on each of the three lines.
7. Find the general solution of the system whose augmented matrix is
\[
\begin{pmatrix}
1 & 2 & 0 & | & 2 \\
1 & 1 & 4 & | & 2 \\
2 & 3 & 4 & | & 4
\end{pmatrix}.
\]

A reduced echelon form for the matrix is
\[
\begin{pmatrix}
1 & 0 & 8 & 2 \\
0 & 1 & -4 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}.
\]

Therefore, \( y = 4z \) and \( x = 2 - 8z \). Apparently \( z \) can equal anything so we let \( z = t \) and then the solution is \( x = 2 - 8t, y = 4t, z = t \).

## A.2 Worked Exercises

1. Here are some matrices:
\[
A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 7 \\ 1 & 0 & 1 \end{pmatrix},
B = \begin{pmatrix} 3 & -1 & 2 \\ -3 & 2 & 1 \end{pmatrix},
\]
\[
C = \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 1 & 1 \end{pmatrix},
D = \begin{pmatrix} -1 & 2 \\ 2 & -3 \end{pmatrix},
E = \begin{pmatrix} 2 \\ 3 \end{pmatrix}.
\]

Find if possible \(-3A, 3B - A, AC, CB, EA, DC^T\). If it is not possible explain why.

\[-3A = -3\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 7 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -3 & -6 & -9 \\ -6 & -9 & -21 \\ -3 & 0 & -3 \end{pmatrix} \]

\(3B - A\) is nonsense because the matrices \( B \) and \( A \) are not of the same size.

\[AC = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 7 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 10 & 7 \\ 18 & 14 \\ 2 & 3 \end{pmatrix} \]

There is no problem here because you are doing \((3 \times 3)(3 \times 2)\).

\[CB = \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 3 & -1 & 2 \\ -3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} -3 & 3 & 4 \\ 6 & -1 & 7 \\ 0 & 1 & 3 \end{pmatrix} \]

There is no problem here because you are doing \((3 \times 2)(2 \times 3)\) and the inside numbers match. \( EA \) is nonsense because it is of the form \((2 \times 1)(3 \times 3)\) so since the inside numbers do not match the matrices are not conformable.

\[DC^T = \begin{pmatrix} -1 & 2 \\ 2 & -3 \end{pmatrix} \begin{pmatrix} 1 & 3 & 1 \\ 2 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 3 & -1 & 1 \\ -4 & 3 & -1 \end{pmatrix} \].
2. Let $A = \begin{pmatrix} 0 & 2 \\ 3 & 4 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 2 \\ 1 & k \end{pmatrix}$. Is it possible to choose $k$ such that $AB = BA$? If so, what should $k$ equal?

We just multiply and see if it can happen.

$AB = \begin{pmatrix} 0 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 1 & k \end{pmatrix} = \begin{pmatrix} 2 & 2k \\ 7 & 6 + 4k \end{pmatrix}$.

On the other hand,

$BA = \begin{pmatrix} 1 & 2 \\ 1 & k \end{pmatrix} \begin{pmatrix} 0 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 6 & 10 \\ 3k & 2 + 4k \end{pmatrix}$.

If these were equal you would need to have $6 = 2$ which is not the case. Therefore, there is no way to choose $k$ such that these two matrices will commute.

3. Let $x = (-1, 0, 3)$ and $y = (3, 1, 2)$. Find $x^T y$.

$x^T y = \begin{pmatrix} -1 \\ 0 \\ 3 \end{pmatrix} \begin{pmatrix} 3 & 1 & 2 \end{pmatrix} = \begin{pmatrix} -3 & -1 & -2 \\ 0 & 0 & 0 \\ 9 & 3 & 6 \end{pmatrix}$.

4. Write \( \begin{pmatrix} 4x_1 - x_2 + 2x_3 \\ 2x_3 + 7x_1 \\ 3x_3 + 3x_2 + x_1 \end{pmatrix} \) in the form $A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$ where $A$ is an appropriate matrix.

\[
\begin{pmatrix} 4 & -1 & 2 & 0 \\ 7 & 0 & 2 & 0 \\ 0 & 0 & 2 & 0 \\ 1 & 3 & 3 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}.
\]

5. Let $A = \begin{pmatrix} 1 & 2 & 5 \\ 2 & 1 & 4 \\ 1 & 0 & 2 \end{pmatrix}$.

Find $A^{-1}$ if possible. If $A^{-1}$ does not exist, determine why.

$\begin{pmatrix} 1 & 2 & 5 \\ 2 & 1 & 4 \\ 1 & 0 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} -2 & 4 & -1 \\ 0 & 1 & -2 \\ 1/3 & -2/3 & 1 \end{pmatrix}$.

6. Let $A = \begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 5 & 2 & 0 \\ 2 & 1 & -3 & 2 \\ 1 & 2 & 1 & 2 \end{pmatrix}$.

Find $A^{-1}$ if possible. If $A^{-1}$ does not exist, determine why.

$\begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 5 & 2 & 0 \\ 2 & 1 & -3 & 2 \\ 1 & 2 & 1 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} -3 & 1/6 & 5/6 & -3/6 \\ -1 & 0 & 0 & 1 \\ 1 & -1/4 & -1/4 & -1/4 \end{pmatrix}$.
7. Show that if $A^{-1}$ exists for an $n \times n$ matrix, then it is unique. That is, if $BA = I$ and $AB = I$, then $B = A^{-1}$.

From $AB = I$, multiply both sides by $A^{-1}$. Thus $A^{-1}(AB) = A^{-1}$. Then from the associative property of matrix multiplication, $A^{-1} = A^{-1}(AB) = (A^{-1}A)B = IB = B$.

**A.3 Worked Exercises**

1. Find the following determinant by expanding along the second column.

$$
\begin{vmatrix}
1 & 3 & 1 \\
2 & 1 & 5 \\
2 & 1 & 1 \\
\end{vmatrix}
$$

This is

$$
3(-1)^{2+1} \begin{vmatrix} 2 & 5 \\ 2 & 1 \end{vmatrix} + 1(-1)^{1+1} \begin{vmatrix} 1 & 1 \\ 2 & 1 \end{vmatrix} + 1(-1)^{3+2} \begin{vmatrix} 1 & 1 \\ 2 & 5 \end{vmatrix} = 20.
$$

2. Compute the determinant by cofactor expansion. Pick the easiest row or column to use.

$$
\begin{vmatrix}
2 & 0 & 0 & 1 \\
2 & 1 & 1 & 0 \\
0 & 0 & 0 & 3 \\
2 & 3 & 3 & 1 \\
\end{vmatrix}
$$

You ought to use the third row. This yields the above equals

$$
3 \begin{vmatrix} 2 & 0 & 0 \\ 2 & 1 & 1 \\ 2 & 3 & 3 \end{vmatrix} = (3)(2) \begin{vmatrix} 1 & 1 \\ 3 & 3 \end{vmatrix} = 0.
$$

3. Find the determinant using row and column operations.

$$
\begin{vmatrix}
5 & 4 & 3 & 2 \\
3 & 2 & 4 & 3 \\
-1 & 2 & 3 & 3 \\
2 & 1 & 2 & -2 \\
\end{vmatrix}
$$

Replace the first row by 5 times the third added to it and then replace the second by 3 times the third added to it and then the last by 2 times the third added to it. This yields

$$
\begin{vmatrix}
0 & 14 & 18 & 17 \\
0 & 8 & 13 & 12 \\
-1 & 2 & 3 & 3 \\
0 & 5 & 8 & 4 \\
\end{vmatrix}
$$

Now let’s replace the third column by −1 times the last column added to it.

$$
\begin{vmatrix}
0 & 14 & 1 & 17 \\
0 & 8 & 1 & 12 \\
-1 & 2 & 0 & 3 \\
0 & 5 & 4 & 4 \\
\end{vmatrix}
$$
Now replace the top row by \(-1\) times the second added to it and the bottom row by \(-4\) times the second added to it. This yields

\[
\begin{vmatrix}
0 & 6 & 0 & 5 \\
0 & 8 & 1 & 12 \\
-1 & 2 & 0 & 3 \\
0 & -27 & 0 & -44
\end{vmatrix}.
\]

(1.1)

This looks pretty good because it has a lot of zeros. Expand along the first column and next along the second,

\[
(-1) \begin{vmatrix} 6 & 0 & 5 \\
8 & 1 & 12 \\
-27 & 0 & -44 \\
\end{vmatrix} = (-1)(1) \begin{vmatrix} 6 & 5 \\
-27 & -44 \\
\end{vmatrix} = 129.
\]

Alternatively, you could continue doing row and column operations. Switch the third and first row in (1.1) to obtain

\[
\begin{vmatrix}
-1 & 2 & 0 & 3 \\
0 & 8 & 1 & 12 \\
0 & 6 & 0 & 5 \\
0 & -27 & 0 & -44
\end{vmatrix}.
\]

Next take \(9/2\) times the third row and add to the bottom.

\[
\begin{vmatrix}
-1 & 2 & 0 & 3 \\
0 & 8 & 1 & 12 \\
0 & 6 & 0 & 5 \\
0 & 0 & 0 & -44 + \left(\frac{9}{2}\right) 5
\end{vmatrix}.
\]

Finally, take \(-6/8\) times the second row and add to the third.

\[
\begin{vmatrix}
-1 & 2 & 0 & 3 \\
0 & 8 & 1 & 12 \\
0 & 0 & -6/8 & 5 + \left(-\frac{6}{8}\right)(12) \\
0 & 0 & 0 & -44 + \left(\frac{9}{2}\right) 5
\end{vmatrix}.
\]

Therefore, since the matrix is now upper triangular, the determinant is

\[
-((-1)(8)(-6/8)(-44 + (9/2)5)) = 129.
\]

4. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.

\[
\begin{pmatrix} a & b \\
 c & d \end{pmatrix} \cdot \begin{pmatrix} a & c \\
 b & d \end{pmatrix}
\]

This involved taking the transpose so the determinant of the new matrix is the same as the determinant of the first matrix.

5. Show that for \(A\) a \(2 \times 2\) matrix \(\det(aA) = a^2 \det(A)\) where \(a\) is a scalar.

\(a^2 \det(A) = a \det(A_1)\) where the first row of \(A\) is replaced by \(a\) times it to get \(A_1\). Then \(a \det(A_1) = A_2\) where \(A_2\) is obtained from \(A\) by multiplying both rows by \(a\). In other words, \(A_2 = aA\). Thus the conclusion is established.
6. Use Cramer’s rule to find \( y \) in
\[
\begin{align*}
2x + 2y + z &= 3 \\
2x - y - z &= 2 \\
x + 2z &= 1
\end{align*}
\]
From Cramer’s rule,
\[
y = \frac{\begin{vmatrix} 2 & 3 & 1 \\ 1 & 1 & 2 \\ 2 & 2 & 1 \end{vmatrix}}{\begin{vmatrix} 2 & 2 & 1 \\ 1 & 1 & 2 \\ 2 & -1 & 1 \end{vmatrix}} = \frac{5}{13}
\]

7. Here is a matrix,
\[
\begin{pmatrix}
e^t & e^{-t} \cos t & e^{-t} \sin t \\
e^t & -e^{-t} \cos t - e^{-t} \sin t & -e^{-t} \sin t + e^{-t} \cos t \\
e^t & 2e^{-t} \sin t & -2e^{-t} \cos t
\end{pmatrix}
\]
Does there exist a value of \( t \) for which this matrix fails to have an inverse? Explain.
\[
\det \begin{pmatrix} e^t & e^{-t} \cos t & e^{-t} \sin t \\ e^t & -e^{-t} \cos t - e^{-t} \sin t & -e^{-t} \sin t + e^{-t} \cos t \\ e^t & 2e^{-t} \sin t & -2e^{-t} \cos t \end{pmatrix} = 5e^t e^{2(-t)} \cos^2 t + 5e^t e^{2(-t)} \sin^2 t
\]
\[
= 5e^{-t} \text{ which is never equal to zero for any value of } t \text{ and so there is no value of } t \text{ for which the matrix has no inverse.}
\]

8. Use the formula for the inverse in terms of the cofactor matrix to find if possible the inverse of the matrix
\[
\begin{pmatrix}
1 & 2 & 3 \\
0 & 6 & 1 \\
4 & 1 & 1
\end{pmatrix}
\]
First you need to take the determinant
\[
\det \begin{pmatrix} 1 & 2 & 3 \\ 0 & 6 & 1 \\ 4 & 1 & 1 \end{pmatrix} = -59
\]
and so the matrix has an inverse. Now you need to find the cofactor matrix.
\[
\begin{pmatrix}
6 & 1 & -0 & 1 & 0 & 6 \\
1 & 1 & 4 & 1 & 4 & 1 \\
-2 & 3 & 1 & 3 & -1 & 2 \\
1 & 1 & 4 & 1 & 4 & 1 \\
2 & 3 & -1 & 3 & 1 & 2 \\
6 & 1 & 0 & 1 & 0 & 6
\end{pmatrix}
\]
\[
= \begin{pmatrix} 5 & 4 & -24 \\ 1 & -11 & 7 \\ -16 & -1 & 6 \end{pmatrix}
\]
Thus the inverse is
\[
\begin{pmatrix}
\frac{1}{-59} \\
\end{pmatrix}
= \begin{pmatrix}
5 & 4 & -24 \\
1 & -11 & 7 \\
-16 & -1 & 6 \\
\end{pmatrix}^T
\]

If you check this, it does work.

## A.4 Worked Exercises

1. Here are three vectors. Determine whether they are linearly independent or linearly dependent.

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

You need to consider the solutions to the equation

\[
c_1 \begin{pmatrix}
1 \\
0 \\
1 \\
\end{pmatrix} + c_2 \begin{pmatrix}
2 \\
0 \\
1 \\
\end{pmatrix} + c_3 \begin{pmatrix}
3 \\
0 \\
0 \\
\end{pmatrix} = \begin{pmatrix}
0 \\
0 \\
0 \\
\end{pmatrix}
\]

and determine whether there is a solution other than the obvious one, \( c_1 = c_2 = c_3 = 0 \).

The augmented matrix for the system of equations is

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

Taking \(-1\) times the top row and adding to the bottom and then switching the two bottom rows yields

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

Next take 2 times the second row and add to the top. This yields

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

There are solutions other than the zero solution because \( c_3 \) is a free variable. Therefore, these vectors are not linearly independent.

2. Verify that the set of continuous functions defined on \([0,1]\) is a vector space.

First of all, the vector space operations are defined on this set of functions because if \( f \) and \( g \) are both continuous, so is their sum. Recall why this is so. Let \( x \in [0,1] \).

Since \( f, g \) are both continuous, it follows there exists \( \delta_1 > 0 \) such that if \( |x - y| < \delta_1 \),
then \(|f(x) - f(y)| < \varepsilon/2\) and there exists \(\delta_2 > 0\) such that if \(|x - y| < \delta_2\), then
\(|g(x) - g(y)| < \varepsilon/2\). Letting \(\delta = \min(\delta_1, \delta_2)\), it follows that if \(|x - y| < \delta\), then
\[
|((f + g)(x) - (f + g)(y))| = |f(x) + g(x) - (f(y) + g(y))| \\
\leq |f(x) - f(y)| + |g(x) - g(y)| \\
< \varepsilon/2 + \varepsilon/2 = \varepsilon.
\]
Since \(\varepsilon\) is arbitrary and \(x\) is arbitrary, this shows \(f + g\) is continuous. Thus the operation of addition of two continuous functions defined in the text by \(f + g)(x) = f(x) + g(x)\) yields a continuous function. What about the operation of scalar multiplication? Suppose \(\alpha\) is a scalar and \(f\) is a continuous function. We need to verify that \(\alpha f\) is a continuous function. Let \(\varepsilon > 0\) be given and suppose \(x \in [0, 1]\). There exists \(\delta > 0\) such that if \(|y - x| < \delta\), then \(|f(x) - f(y)| < \varepsilon/(1 + |\alpha|)\). Then if \(|x - y| < \delta\),
\[
|\alpha f(x) - \alpha f(y)| \leq |\alpha||f(x) - f(y)| \\
\leq |\alpha|\frac{\varepsilon}{1 + |\alpha|} < \varepsilon
\]
and so \(\alpha f\) is continuous at each \(x \in [0, 1]\). This has shown the two operations of addition and scalar multiplication are well defined on the set of continuous functions defined on \([0, 1]\). What about the vector space axioms?

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

Since \((f + g)(x) = (g + f)(x)\) for every \(x\), it follows \(f + g = g + f\). This proves the commutative law of vector addition. Next

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

and since \(x\) is arbitrary, \((f + g) + h = f + (g + h)\) so the associative law holds. The zero function, that function defined by \(0(x) = 0\) for all \(x\) is an additive identity because \((0 + f)(x) \equiv 0(x) + f(x) = f(x)\) and so \(0 + f = f\) because \(x\) is arbitrary. There exists an additive inverse given by \((-1)f\) because

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

and since \(x\) is arbitrary, \((-1)f\) is an additive inverse for \(f\). What about the distributive laws?

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

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

\[
((\alpha + \beta)f)(x) \equiv (\alpha + \beta)f(x) \\
= \alpha f(x) + \beta f(x) \\
= (\alpha f + \beta f)(x)
\]

and since \(x\) is arbitrary, \((\alpha + \beta)f = \alpha f + \beta f\). What about the associative law for multiplication?

\[
(\alpha (\beta f))(x) \equiv \alpha (\beta f)(x) \equiv \alpha \beta f(x) = ((\alpha \beta) f)(x)
\]
and so $\alpha (\beta f) = (\alpha \beta) f$ since $x$ was arbitrary. Finally, $(1f)(x) \equiv 1(f(x)) = f(x)$ and since $x$ is arbitrary, $1f = f$. This verifies the vector space axioms for functions defined on a domain $D$ in the special case where $D = [0, 1]$. We leave the generalization of this for arbitrary $D$ to you.

3. Here are four vectors. Determine whether they span $\mathbb{R}^3$. Are these vectors linearly independent?

\[
\begin{pmatrix}
1 \\
2 \\
3
\end{pmatrix},
\begin{pmatrix}
4 \\
0 \\
3
\end{pmatrix},
\begin{pmatrix}
3 \\
2 \\
0
\end{pmatrix},
\begin{pmatrix}
2 \\
1 \\
6
\end{pmatrix}
\]

The vectors can't possibly be linearly independent. If they were, they would constitute a linearly independent set consisting of four vectors even though there exists a spanning set of only three,

\[
\begin{pmatrix}
1 \\
0 \\
0
\end{pmatrix},
\begin{pmatrix}
0 \\
1 \\
0
\end{pmatrix},
\begin{pmatrix}
0 \\
0 \\
1
\end{pmatrix}
\]

contrary to the conclusion of the exchange theorem. However, the four given vectors might still span $\mathbb{R}^3$ even though they are not a basis. What does it take to span $\mathbb{R}^3$? Given a vector $(x, y, z)^T \in \mathbb{R}^3$, do there exist scalars $c_1, c_2, c_3,$ and $c_4$ such that

\[
c_1 \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + c_2 \begin{pmatrix} 4 \\ 0 \\ 3 \end{pmatrix} + c_3 \begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix} + c_4 \begin{pmatrix} 2 \\ 1 \\ 6 \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}?
\]

Consider the augmented matrix of the above,

\[
\begin{pmatrix}
1 & 4 & 3 & 2 & | & x \\
2 & 0 & 2 & 1 & | & y \\
3 & 3 & 0 & 6 & | & z
\end{pmatrix}
\]

Doing row operations till an echelon form is obtained leads to

\[
\begin{pmatrix}
1 & 0 & 0 & \frac{5}{4} & | & \frac{1}{4}y + \frac{5}{3}z - \frac{1}{6}x \\
0 & 1 & 0 & \frac{1}{4} & | & -\frac{1}{2}y + \frac{1}{3}x + \frac{1}{6}z \\
0 & 0 & 1 & -\frac{3}{4} & | & -\frac{5}{3}z + \frac{1}{3}x + \frac{1}{6}y
\end{pmatrix}
\]

and you see there is a solution to the desired system of equations. In fact there are infinitely many because $c_4$ is a free variable. Therefore, the four vectors do span $\mathbb{R}^3$.

4. Consider the vectors of the form

\[
\begin{pmatrix}
2t + 6s \\
s - 2t \\
3t + s
\end{pmatrix} : s, t \in \mathbb{R}
\]

Is this set of vectors a subspace of $\mathbb{R}^3$? If so, explain why, give a basis for the subspace and find its dimension.

This is indeed a subspace. You only need to verify the set of vectors is closed with respect to the vector space operations. Let \( \begin{pmatrix} 2t_1 + 6s_1 \\ s_1 - 2t_1 \\ 3t_1 + s_1 \end{pmatrix} \) and \( \begin{pmatrix} 2t + 6s \\ s - 2t \\ 3t + s \end{pmatrix} \) be two
vectors in the given set of vectors.

\[
\begin{align*}
\alpha \begin{pmatrix} 2t + 6s \\ s - 2t \\ 3t + s \end{pmatrix} + \beta \begin{pmatrix} 2t_1 + 6s_1 \\ s_1 - 2t_1 \\ 3t_1 + s_1 \end{pmatrix} &= \begin{pmatrix} 2\alpha t + 6\alpha s + 2\beta t_1 + 6\beta s_1 \\ \alpha s - 2\alpha t + \beta s_1 - 2\beta t_1 \\ 3\alpha t + \alpha s + 3\beta t_1 + \beta s_1 \end{pmatrix} \\
&= \begin{pmatrix} 2(\alpha t + \beta t_1) + 6(\alpha s + \beta s_1) \\ \alpha s + \beta s_1 - 2(\alpha t + \beta t_1) \\ 3(\alpha t + \beta t_1) + \alpha s + \beta s_1 \end{pmatrix}
\end{align*}
\]

If we let \( T \equiv \alpha t + \beta t_1 \) and \( S \equiv \alpha s + \beta s_1 \), this is seen to be of the form

\[
\begin{pmatrix} 2T + 6S \\ S - 2T \\ 3T + S \end{pmatrix}
\]

which is the way the vectors in the given set are described. Another way to see this is to notice that the vectors in the given set are of the form

\[
\begin{pmatrix} 2t \\ -2 \\ 3 \end{pmatrix} + s \begin{pmatrix} 6 \\ 1 \\ 1 \end{pmatrix}
\]

so it consists of the span of the two vectors,

\[
\begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 6 \\ 1 \\ 1 \end{pmatrix}
\]

Recall that the span of a set of vectors is always a subspace. You can also verify these vectors in (1.2) form a linearly independent set and so they are a basis.

5. Let \( M = \{ \mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_3 \geq u_2 \} \). Is \( M \) a subspace? Explain.

This is not a subspace because if \( \mathbf{u} \in M \), is such that \( u_3 > u_2 \), then consider \((-1) \mathbf{u}\).

If this were in \( M \) you would need to have \(-u_3 > -u_2\) and so \( u_3 < u_2 \) which cannot be true if \( u_3 > u_2 \). Thus \( M \) is not closed under scalar multiplication so it is not a subspace.

6. Let \( \mathbf{w}, \mathbf{w}_1 \) be given vectors in \( \mathbb{R}^2 \) and define

\[
M = \{ \mathbf{u} = (u_1, u_2) \in \mathbb{R}^2 : \mathbf{w} \cdot \mathbf{u} = 0 \text{ and } \mathbf{w}_1 \cdot \mathbf{u} = 0 \}.
\]

Is \( M \) a subspace? Explain.

Suppose \( \mathbf{u}' \) and \( \mathbf{u} \) are both in \( M \). What about \( \alpha \mathbf{u}' + \beta \mathbf{u} \)?

\[
\mathbf{w} \cdot (\alpha \mathbf{u}' + \beta \mathbf{u}) = \alpha \mathbf{w} \cdot \mathbf{u}' + \beta \mathbf{w} \cdot \mathbf{u} = \alpha 0 + \beta 0 = 0
\]

Similarly,

\[
\mathbf{w}_1 \cdot (\alpha \mathbf{u}' + \beta \mathbf{u}) = \alpha \mathbf{w}_1 \cdot \mathbf{u}' + \beta \mathbf{w}_1 \cdot \mathbf{u} = \alpha 0 + \beta 0 = 0
\]

and so \( \alpha \mathbf{u}' + \beta \mathbf{u} \in M \). This has verified that \( M \) is a subspace.
7. In any vector space, show that if \( \mathbf{x} + \mathbf{y} = \mathbf{0} \), then \( \mathbf{y} = -\mathbf{x} \).

\[
-\mathbf{x} = -\mathbf{x} + (\mathbf{x} + \mathbf{y}) = (-\mathbf{x} + \mathbf{x}) + \mathbf{y} = \mathbf{0} + \mathbf{y} = \mathbf{y}.
\]

8. Show that in any vector space, \( 0 \mathbf{x} = \mathbf{0} \). That is, the scalar 0 times the vector \( \mathbf{x} \) gives the vector \( \mathbf{0} \).

\[
0 \mathbf{x} = (0 + 0) \mathbf{x} = 0 \mathbf{x} + 0 \mathbf{x}.
\]

Now add \( -(0 \mathbf{x}) \) to both sides to conclude

\[
0 = -(0 \mathbf{x}) + (0 \mathbf{x}) = -(0 \mathbf{x} + 0 \mathbf{x})
\]

\[
(-0 \mathbf{x}) + 0 \mathbf{x} = 0 + 0 \mathbf{x} = 0 \mathbf{x}.
\]

Therefore, \( 0 \mathbf{x} = \mathbf{0} \) as claimed.

9. Show that in any vector space, \( (-1) \mathbf{x} = -\mathbf{x} \).

From Problem 7 all we need to do is show that \( (-1) \mathbf{x} \) acts like \( -\mathbf{x} \) and then it follows that it is \( -\mathbf{x} \).

\[
(-1) \mathbf{x} + \mathbf{x} = ((-1) + 1) \mathbf{x} = 0 \mathbf{x} = \mathbf{0},
\]

the last equal sign holding because of Problem 8. Therefore, from Problem 7, it follows \( -\mathbf{x} = (-1) \mathbf{x} \).

A.5 Worked Exercises

1. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( 5\pi/12 \).

You note that \( 5\pi/12 = 2\pi/3 - \pi/4 \). Therefore, you can first rotate through \( -\pi/4 \) and then rotate through \( 2\pi/3 \) to get the rotation through \( 5\pi/12 \). The matrix of the transformation with respect to the usual coordinates which rotates through \( -\pi/4 \) is

\[
\begin{pmatrix}
\sqrt{2}/2 & \sqrt{2}/2 \\
-\sqrt{2}/2 & \sqrt{2}/2
\end{pmatrix}
\]

and the matrix of the transformation which rotates through \( 2\pi/3 \) is

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

Multiplying these gives

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

\[
= \begin{pmatrix}
-\frac{1}{4}\sqrt{2} + \frac{1}{4}\sqrt{3}\sqrt{2} & -\frac{1}{4}\sqrt{2} - \frac{1}{4}\sqrt{3}\sqrt{2} \\
\frac{1}{4}\sqrt{3}\sqrt{2} + \frac{1}{4}\sqrt{2} & -\frac{1}{4}\sqrt{2} + \frac{1}{4}\sqrt{3}\sqrt{2}
\end{pmatrix}
\]

and this is the matrix of the desired transformation. Note this shows that

\[
\cos (5\pi/12) = -\frac{1}{4}\sqrt{2} + \frac{1}{4}\sqrt{3}\sqrt{2} \approx 0.258 819 05
\]

\[
\sin (5\pi/12) = \frac{1}{4}\sqrt{3}\sqrt{2} + \frac{1}{4}\sqrt{2} \approx 0.965 925 83.
\]
2. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( 2 \pi/3 \) and then reflects across the \( x \) axis.

What does it do to \( e_1 \)? First you rotate \( e_1 \) through the given angle to obtain

\[
\begin{pmatrix}
-1/2 \\
\sqrt{3}/2
\end{pmatrix}
\]

and then this becomes

\[
\begin{pmatrix}
-1/2 \\
-\sqrt{3}/2
\end{pmatrix}.
\]

This is the first column of the desired matrix. Next \( e_2 \) first is rotated through the given angle to give

\[
\begin{pmatrix}
-\sqrt{3}/2 \\
-1/2
\end{pmatrix}
\]

and then it is reflected across the \( x \) axis to give

\[
\begin{pmatrix}
-\sqrt{3}/2 \\
1/2
\end{pmatrix}
\]

and this gives the second column of the desired matrix. Thus the matrix is

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

3. Find the matrix for \( \text{proj}_u (v) \) where \( u = (1, -2, 3)^T \).

Recall

\[
\text{proj}_u (v) = \frac{v \cdot u}{||u||^2} u
\]

Therefore,

\[
\text{proj}_u (e_1) = \frac{1}{14} \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix}, \quad \text{proj}_u (e_2) = \frac{-2}{14} \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix},
\]

\[
\text{proj}_u (e_2) = \frac{3}{14} \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix}.
\]

Hence the desired matrix is

\[
\frac{1}{14} \begin{pmatrix} 1 & -2 & 3 \\ -2 & 4 & -6 \\ 3 & -6 & 9 \end{pmatrix}.
\]

4. Show that the function \( T_u \) defined by \( T_u (v) \equiv v - \text{proj}_u (v) \) is also a linear transformation.

\[
T_u (\alpha v + \beta w) = \alpha v + \beta w - \text{proj}_u (\alpha v + \beta w)
\]

which from 3 equals

\[
\alpha (v - \text{proj}_u (v)) + \beta (w - \text{proj}_u (w)) = \alpha T_u v + \beta T_u w.
\]

This is what it takes to be a linear transformation.
5. If $A$, $B$, and $C$ are each $n \times n$ matrices and $ABC$ is invertible, why are each of $A$, $B$, and $C$ invertible.

$0 \neq \det(ABC) = \det(A) \det(B) \det(C)$ and so none of $\det(A)$, $\det(B)$, or $\det(C)$ can equal zero. Therefore, each is invertible. You should do this another way, showing that each of $A$, $B$, and $C$ is one to one and then using a theorem presented earlier.

6. Give an example of a $3 \times 1$ matrix with the property that the linear transformation determined by this matrix is one to one but not onto.

Here is one. 

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

If 

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} x = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

then $x = 0$ but this is certainly not onto as a map from $\mathbb{R}^1$ to $\mathbb{R}^3$ because it does not ever yield 

$$\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

7. Find the matrix of the linear transformation from $\mathbb{R}^3$ to $\mathbb{R}^3$ which first rotates every vector through an angle of $\pi/4$ about the $z$ axis when viewed from the positive $z$ axis and then rotates every vector through an angle of $\pi/6$ about the $x$ axis when viewed from the positive $x$ axis.

The matrix of the linear transformation which accomplishes the first rotation is

$$\begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 & 0 \\ \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and the matrix which accomplishes the second rotation is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \sqrt{3}/2 & -1/2 \\ 0 & 1/2 & \sqrt{3}/2 \end{pmatrix}$$

Therefore, the matrix of the desired linear transformation is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \sqrt{3}/2 & -1/2 \\ 0 & 1/2 & \sqrt{3}/2 \end{pmatrix} \begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 & 0 \\ \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1/2\sqrt{2} & -1/2\sqrt{2} & 0 \\ 1/4\sqrt{3}\sqrt{2} & 1/4\sqrt{3}\sqrt{2} & -1/2 \\ 1/4\sqrt{2} & 1/4\sqrt{3} & 1/2 \end{pmatrix}$$

This might not be the first thing you would think of.

8. In the study of rotating bodies, more complex situations are considered. These problems are studied through the use of Euler angles. To describe the Euler angles we consider the following picture in which $x_1$, $x_2$ and $x_3$ are the usual coordinate axes fixed in space and the axes labeled with a superscript denote other coordinate axes. Here is the picture.
We obtain $\phi$ by rotating about the fixed $x_3$ axis. Next we rotate about the $x_1^1$ axis which results from the first rotation. This gives $\theta$. Finally, we rotate about the $x_2^3$ axis by $\psi$. One finds differential equations for the motion of spinning bodies through the use of these angles by writing a formula for the kinetic energy in terms of the angles shown and then using a formalism developed by Lagrange back in the 1700’s to obtain the differential equations. To
see how these angles relate to a spinning object, consider the following picture.

A.6 Worked Exercises

1. Find the rank of the following matrices. If the rank is $r$, identify $r$ columns in the original matrix which have the property that every other column may be written as a linear combination of these. Also find a basis for the row and column spaces of the matrices.

(a) \[
\begin{pmatrix}
9 & 2 & 0 \\
3 & 7 & 1 \\
6 & 1 & 0 \\
0 & 2 & 1 \\
\end{pmatrix}
\]

From using row operations we obtain the row reduced echelon form which is

\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 0 \\
\end{pmatrix}
\]
Therefore, a basis for the column space of the original matrix is the first three columns of the original matrix. A basis for the row space is just \((1\ 0\ 0)\), \((0\ 1\ 0)\), and \((0\ 0\ 1)\).

(b) \[
\begin{pmatrix}
3 & 0 & 3 \\
10 & 9 & 1 \\
1 & 1 & 0 \\
2 & 2 & 0
\end{pmatrix}
\]
In this case the row reduced echelon form is
\[
\begin{pmatrix}
1 & 0 & 1 \\
0 & 1 & -1 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
\]
and so a basis for the column space of the original matrix consists of the first two columns of the original matrix and a basis for the row space is \((1\ 0\ 1)\) and \((0\ 1\ -1)\).

(c) \[
\begin{pmatrix}
0 & 1 & 7 & 8 & 1 & 9 & 2 \\
0 & 3 & 2 & 5 & 1 & 6 & 8 \\
0 & 1 & 1 & 2 & 0 & 2 & 3 \\
0 & 2 & 1 & 3 & 0 & 3 & 4
\end{pmatrix}
\]
The row reduced echelon form of this matrix is
\[
\begin{pmatrix}
0 & 1 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
\]
and so a basis for the column space of the original matrix consists of the second, third, fifth, and seventh columns of the original matrix. A basis for the row space consists of the rows of this last matrix in row reduced echelon form.

2. Let \(H\) denote \(\text{span}\left\{\begin{pmatrix}1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix}1 \\ 4 \\ 5 \end{pmatrix}, \begin{pmatrix}1 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix}1 \\ 1 \\ 1 \end{pmatrix}\right\}\). Find the dimension of \(H\) and determine a basis.

Make these the columns of a matrix and ask for the rank of this matrix.
\[
\begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 4 & 3 & 1 \\
0 & 5 & 1 & 1
\end{pmatrix}
\]
The row reduced echelon form is
\[
\begin{pmatrix}
1 & 0 & 0 & \frac{8}{7} \\
0 & 1 & 0 & \frac{4}{7} \\
0 & 0 & 1 & -\frac{2}{7}
\end{pmatrix}
\]
A basis for \(H\) is
\[
\left\{\begin{pmatrix}1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix}1 \\ 4 \\ 5 \end{pmatrix}, \begin{pmatrix}1 \\ 3 \\ 1 \end{pmatrix}\right\}
\]
and so \(H\) has dimension 3.
### A.7 Worked Exercises

1. Find an $LU$ decomposition of \( \begin{pmatrix} 1 & 2 & 7 \\ 3 & 1 & 3 \\ 1 & 2 & 3 \end{pmatrix} \).

To find this we write
\[
\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 7 \\ 3 & 1 & 3 \\ 1 & 2 & 3 \end{pmatrix}
\]
and put the one on the right into echelon form and keep track of the multipliers. Updating the first column,
\[
\begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 7 \\ 0 & -5 & -18 \\ 0 & 0 & -4 \end{pmatrix}
\]
We now stop because the matrix on the right is upper triangular.

2. Find an $LU$ decomposition of \( \begin{pmatrix} 1 & 7 & 3 & 2 \\ 1 & 3 & 8 & 1 \\ 5 & 1 & 1 & 3 \end{pmatrix} \).

To find it we write
\[
\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 7 & 3 & 2 \\ 1 & 3 & 8 & 1 \\ 5 & 1 & 1 & 3 \end{pmatrix}
\]
and update keeping track of the multipliers. First we update the first column.
\[
\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 5 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 7 & 3 & 2 \\ 0 & -4 & 5 & -1 \\ 0 & -34 & -14 & -7 \end{pmatrix}
\]
Next we update the second column.
\[
\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 5 & 34/4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 7 & 3 & 2 \\ 0 & -4 & 5 & -1 \\ 0 & -113/2 & 3/2 \end{pmatrix}
\]
At this point we stop because the matrix on the right is in upper triangular form.

3. Find the $LU$ decomposition of the coefficient matrix using Dolittle’s method and use it to solve the system of equations.

\[
x + 2y + 3z = 5 \\
2x + 3y + 3z = 6 \\
3x + 5y + 4z = 11
\]

The coefficient matrix is
\[
\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 3 \\ 3 & 5 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -1 & -3 \\ 0 & 0 & -2 \end{pmatrix}.
\]
Then we first solve
\[
\begin{pmatrix}
1 & 0 & 0 \\
2 & 1 & 0 \\
3 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
u \\
v \\
w
\end{pmatrix}
=
\begin{pmatrix}
5 \\
6 \\
11
\end{pmatrix}
\]
which yields
\[
\begin{pmatrix}
u \\
v \\
w
\end{pmatrix}
=
\begin{pmatrix}
5 \\
-4 \\
0
\end{pmatrix}
\]
Next we solve
\[
=\left(u, v, w\right)^T
\]
\[
\begin{pmatrix}
1 & 2 & 3 \\
0 & -1 & -3 \\
0 & 0 & -2
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
=
\begin{pmatrix}
5 \\
-4 \\
0
\end{pmatrix}
\]
which yields
\[
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
=
\begin{pmatrix}
-3 \\
4 \\
0
\end{pmatrix}
\]

A.8 Worked Exercises

1. Let \(M\) be an \(n \times n\) matrix. Then define the adjoint of \(M\), denoted by \(M^*\) to be the transpose of the conjugate of \(M\). For example,
\[
\begin{pmatrix}
2 & i \\
1 + i & 3
\end{pmatrix}^*
=\begin{pmatrix}
2 & 1 - i \\
-i & 3
\end{pmatrix}.
\]
A matrix, \(M\), is self adjoint if \(M^* = M\). Show the eigenvalues of a self adjoint matrix are all real. If the self adjoint matrix has all real entries, it is called symmetric. Show that the eigenvalues and eigenvectors of a symmetric matrix occur in conjugate pairs.
First note that for \(x\) a vector, \(x^*x = |x|^2\). This is because
\[
x^*x = \sum_k \bar{x_k}x_k = \sum_k |x_k|^2 \equiv |x|^2.
\]
Also note that \((AB)^* = B^*A^*\) because this holds for transposes. This implies that for \(A\) an \(n \times m\) matrix,
\[
x^*A^*x = (Ax)^*x
\]
Then if \(Mx = \lambda x\)
\[
\bar{x}^*x = (\lambda x)^*x = (Mx)^*x = x^*M^*x
\]
and so \(\lambda = \bar{\lambda}\) showing that \(\lambda\) must be real.

2. Suppose \(A\) is an \(n \times n\) matrix consisting entirely of real entries but \(a + ib\) is a complex eigenvalue having the eigenvector, \(x + iy\). Here \(x\) and \(y\) are real vectors. Show that then \(a - ib\) is also an eigenvalue with the eigenvector, \(x - iy\). Hint: You should remember that the conjugate of a product of complex numbers equals the product of the conjugates. Here \(a + ib\) is a complex number whose conjugate equals \(a - ib\).
If $A$ is real then the characteristic equation has all real coefficients. Therefore, letting $p(\lambda)$ be the characteristic polynomial,

$$0 = p(\lambda) = \overline{p(\lambda)} = p(\bar{\lambda})$$

showing that $\bar{\lambda}$ is also an eigenvalue.

3. Find the eigenvalues and eigenvectors of the matrix

$$
\begin{pmatrix}
-10 & -2 & 11 \\
-18 & 6 & -9 \\
10 & -10 & -2
\end{pmatrix}.
$$

Determine whether the matrix is defective.

The matrix has eigenvalues $-12$ and $18$. Of these, $-12$ is repeated with multiplicity two. Therefore, you need to see whether the eigenspace has dimension two. If it does, then the matrix is non defective. If it does not, then the matrix is defective. The row reduced echelon form for the system you need to solve is

$$
\begin{pmatrix}
2 & -2 & 11 & | & 0 \\
-18 & 18 & -9 & | & 0 \\
10 & -10 & 10 & | & 0
\end{pmatrix}
$$

and its row reduced echelon form is

$$
\begin{pmatrix}
1 & -1 & 0 & | & 0 \\
0 & 0 & 1 & | & 0 \\
0 & 0 & 0 & | & 0
\end{pmatrix}
$$

Therefore, the eigenspace is of the form

$$
\begin{pmatrix}
t \\
t \\
0
\end{pmatrix}
$$

This is only one dimensional and so the matrix is defective.

4. Find the complex eigenvalues and eigenvectors of the matrix

$$
\begin{pmatrix}
1 & 1 & -6 \\
7 & -5 & -6 \\
-1 & 7 & 2
\end{pmatrix}.
$$

Determine whether the matrix is defective.

After wading through much affliction you find the eigenvalues are $-6, 2 + 6i, 2 - 6i$. Since these are distinct, the matrix cannot be defective. We must find the eigenvectors for these eigenvalues. The augmented matrix for the system of equations which must be solved to find the eigenvectors associated with $2 - 6i$ is

$$
\begin{pmatrix}
-1 + 6i & 1 & -6 & | & 0 \\
7 & -7 + 6i & -6 & | & 0 \\
-1 & 7 & 6i & | & 0
\end{pmatrix}.
$$

The row reduced echelon form is

$$
\begin{pmatrix}
1 & 0 & i & 0 \\
0 & 1 & i & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}$$
and so the eigenvectors are of the form
\[
t \begin{pmatrix} -i \\ -i \\ 1 \end{pmatrix}.
\]

You can check this as follows
\[
\begin{pmatrix} 1 & 1 & -6 \\ 7 & -5 & -6 \\ -1 & 7 & 2 \end{pmatrix} \begin{pmatrix} -i \\ -i \\ 1 \end{pmatrix} = \begin{pmatrix} -6 - 2i \\ -6 - 2i \\ 2 - 6i \end{pmatrix}
\]
and
\[
(2 - 6i) \begin{pmatrix} -i \\ -i \\ 1 \end{pmatrix} = \begin{pmatrix} -6 - 2i \\ -6 - 2i \\ 2 - 6i \end{pmatrix}.
\]

It follows that the eigenvectors for \( \lambda = 2 + 6i \) are
\[
t \begin{pmatrix} i \\ i \\ 1 \end{pmatrix}.
\]

This is because \( A \) is real. If \( A \mathbf{v} = \lambda \mathbf{v} \), then taking the conjugate,
\[
A \mathbf{v} = \bar{A} \mathbf{v} = \bar{\lambda} \mathbf{v}.
\]

It only remains to find the eigenvector for \( \lambda = -6 \). The augmented matrix to row reduce is
\[
\begin{pmatrix} 7 & 1 & -6 & | & 0 \\ 7 & 1 & -6 & | & 0 \\ -1 & 7 & 8 & | & 0 \end{pmatrix}
\]
The row reduced echelon form is
\[
\begin{pmatrix} 1 & 0 & -1 & | & 0 \\ 0 & 1 & 1 & | & 0 \\ 0 & 0 & 0 & | & 0 \end{pmatrix}.
\]
Then an eigenvector is
\[
\begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}.
\]

5. You own a trailer rental company in a large city and you have four locations, one in the South East, one in the North East, one in the North West, and one in the South West. Denote these locations by SE, NE, NW, and SW respectively. Suppose you observe that in a typical day, .7 of the trailers starting in SE stay in SE, .1 of the trailers in NE go to SE, .1 of the trailers in NW end up in SE, .2 of the trailers in SW end up in SE, .1 of the trailers in NE end up in NW, .7 of the trailers in NW end up in SE, .1 of the trailers in SW end up in SW, .2 of the trailers in NW end up in NE, .6 of the trailers in NW end up in SW, .1 of the trailers in NW end up in SW, .1 of the trailers in SW end up in NW, .5 of the trailers in SW end up in SW. You begin with 20 trailers in
each location. Approximately how many will you have in each location after a long
time? Will any location ever run out of trailers?

It sometimes helps to write down a table summarizing the given information.

<table>
<thead>
<tr>
<th></th>
<th>SE</th>
<th>NE</th>
<th>NW</th>
<th>SW</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE</td>
<td>.7</td>
<td>.1</td>
<td>.1</td>
<td>.2</td>
</tr>
<tr>
<td>NE</td>
<td>.1</td>
<td>.7</td>
<td>.2</td>
<td>.1</td>
</tr>
<tr>
<td>NW</td>
<td>.2</td>
<td>.1</td>
<td>.6</td>
<td>.2</td>
</tr>
<tr>
<td>SW</td>
<td>0</td>
<td>.1</td>
<td>.1</td>
<td>.5</td>
</tr>
</tbody>
</table>

Then the migration matrix is

\[
\begin{pmatrix}
\frac{7}{10} & 1/10 & 1/10 & 1/5 \\
1/10 & \frac{7}{10} & 1/5 & 1/10 \\
1/5 & 1/10 & 3/5 & 1/5 \\
0 & 1/10 & 1/10 & 1/2
\end{pmatrix}
\]

All we have to do is find the eigenvector (In this case the eigenspace will be one
dimensional because some power of the matrix has all positive entries.) corresponding
to \( \lambda = 1 \) which has all the entries add to 20. This will be the long time population.
Remember, these processes conserve the sum of the entries. We must row reduce

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

The row reduced echelon form is

\[
\begin{pmatrix}
1 & 0 & 0 & -\frac{7}{3} & 0 \\
0 & 1 & 0 & -\frac{7}{3} & 0 \\
0 & 0 & 1 & -\frac{3}{2} & 0 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

Therefore, the eigenvectors are of the form

\[
t \begin{pmatrix}
7 \\
8 \\
7 \\
3
\end{pmatrix}
\]

and we simply need to choose \( t \) in such a way that the entries add to 20. Thus

\[7t + 8t + 7t + 3t = 20\]

so \( t = 4/5 \). Then the long time limit equals

\[
\frac{4}{5} \begin{pmatrix}
7 \\
8 \\
7 \\
3
\end{pmatrix} = \begin{pmatrix}
5.6 \\
6.4 \\
5.6 \\
2.4
\end{pmatrix}
\]

Thus there will be about 5.6 trailers in SE, 6.4 in NE, 5.6 in NW, and 2.4 in SW. In
particular, it appears no location will run out of trailers.
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 best proofs of it depend on either analysis or topology. It was first proved by Gauss in 1797. The proof given here follows Rudin [22]. See also Hardy [14] for a similar proof, more discussion and references. The best proof is found in the theory of complex analysis.

Recall De Moivre’s theorem from trigonometry which is listed here for convenience.

**Theorem B.0.1** Let \( r > 0 \) be given. Then if \( n \) is a positive integer,

\[
[r \,(\cos t + i \sin t)]^n = r^n \,(\cos nt + i \sin nt).
\]

Recall that this theorem is the basis for proving the following corollary from trigonometry, also listed here for convenience.

**Corollary B.0.2** Let \( z \) be a non-zero complex number and let \( k \) be a positive integer. Then there are always exactly \( k \) \( k \)th roots of \( z \) in \( \mathbb{C} \).

**Lemma B.0.3** Let \( a_k \in \mathbb{C} \) for \( k = 1, \cdots, n \) and let \( p(z) = \sum_{k=1}^{n} a_k z^k \). Then \( p \) is continuous.

**Proof:**

\[
|az^n - aw^n| \leq |a| |z - w| |z^{n-1} + z^{n-2}w + \cdots + w^{n-1}|.
\]

Then for \( |z - w| < 1 \), the triangle inequality implies \( |w| < 1 + |z| \) and so if \( |z - w| < 1 \),

\[
|az^n - aw^n| \leq |a| |z - w| n (1 + |z|)^n.
\]

If \( \varepsilon > 0 \) is given, let

\[
\delta < \min \left( 1, \frac{\varepsilon}{|a| n (1 + |z|)^n} \right).
\]

It follows from the above inequality that for \( |z - w| < \delta \), \( |az^n - aw^n| < \varepsilon \). The function of the lemma is just the sum of functions of this sort and so it follows that it is also continuous.

**Theorem B.0.4** (Fundamental theorem of Algebra) Let \( p(z) \) be a nonconstant polynomial. Then there exists \( z \in \mathbb{C} \) such that \( p(z) = 0 \).
Proof: Suppose not. Then

\[ p(z) = \sum_{k=0}^{n} a_k z^k \]

where \( a_n \neq 0, \ n > 0 \). Then

\[ |p(z)| \geq |a_n||z|^n - \sum_{k=0}^{n-1} |a_k||z|^k \]

and so

\[ \lim_{|z| \to \infty} |p(z)| = \infty. \quad (2.1) \]

Now let

\[ \lambda \equiv \inf \{ |p(z)| : z \in \mathbb{C} \}. \]

By (2.1), there exists an \( R > 0 \) such that if \(|z| > R\), it follows that \( |p(z)| > \lambda + 1 \). Therefore,

\[ \lambda \equiv \inf \{ |p(z)| : z \in \mathbb{C} \} = \inf \{ |p(z)| : |z| \leq R \}. \]

The set \( \{ z : |z| \leq R \} \) is a closed and bounded set and so this infimum is achieved at some point \( w \) with \(|w| \leq R\). A contradiction is obtained if \( |p(w)| = 0 \) so assume \( |p(w)| > 0 \). Then consider

\[ q(z) \equiv \frac{p(z+w)}{p(w)}. \]

It follows \( q(z) \) is of the form

\[ q(z) = 1 + c_k z^k + \cdots + c_n z^n \]

where \( c_k \neq 0 \), because \( q(0) = 1 \). It is also true that \( |q(z)| \geq 1 \) by the assumption that \( |p(w)| \) is the smallest value of \( |p(z)| \). Now let \( \theta \in \mathbb{C} \) be a complex number with \(|\theta| = 1\) and

\[ \theta c_k w^k = -|w|^k |c_k|. \]

If

\[ w \neq 0, \ \theta = -\frac{|w|^k |c_k|}{w^k c_k} \]

and if \( w = 0, \ \theta = 1 \) will work. Now let \( \eta^k = \theta \) and let \( t \) be a small positive number.

\[ q(t \eta w) \equiv 1 - t^k |w|^k |c_k| + \cdots + c_n t^n (\eta w)^n \]

which is of the form

\[ 1 - t^k |w|^k |c_k| + t^k (g(t,w)) \]

where \( \lim_{t \to 0} g(t,w) = 0 \). Letting \( t \) be small enough,

\[ |g(t,w)| < |w|^k |c_k|/2 \]

and so for such \( t \),

\[ |q(t \eta w)| < 1 - t^k |w|^k |c_k| + t^k |w|^k |c_k|/2 < 1, \]

a contradiction to \( |q(z)| \geq 1 \). This proves the theorem.
C.1 Vector Spaces

It is time to consider the idea of a vector space.

**Definition C.1.1** 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”, \( F \) which are allowed to multiply the vectors according to the following rules. (The Greek letters denote scalars.)

\[
\alpha (v + w) = \alpha v + \alpha w, \quad (3.1)
\]

\[
(\alpha + \beta)v = \alpha v + \beta v, \quad (3.2)
\]

\[
\alpha(\beta v) = \alpha \beta (v), \quad (3.3)
\]

\[
1v = v. \quad (3.4)
\]

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.

For example, \( \mathbb{R}^n \) with the usual conventions is an example of a real vector space and \( \mathbb{C}^n \) is an example of a complex vector space. Up to now, the discussion has been for \( \mathbb{R}^n \) or \( \mathbb{C}^n \) and all that is taking place is an increase in generality and abstraction.

**Definition C.1.2** If \( \{v_1, \ldots, v_n\} \subseteq V \), a vector space, then

\[
\text{span}(v_1, \ldots, v_n) \equiv \left\{ \sum_{i=1}^{n} \alpha_i v_i : \alpha_i \in F \right\}.
\]

A subset, \( W \subseteq V \) is said to be a subspace if it is also a vector space with the same field of scalars. Thus \( W \subseteq V \) is a subspace if \( ax + by \in W \) whenever \( a, b \in F \) and \( x, y \in W \). The span of a set of vectors as just described is an example of a subspace.
Definition C.1.3 If \( \{v_1, \ldots, v_n\} \subseteq V \), the set of vectors is linearly independent if
\[ \sum_{i=1}^{n} \alpha_i v_i = 0 \]
implies
\[ \alpha_1 = \cdots = \alpha_n = 0 \]
and \( \{v_1, \ldots, v_n\} \) is called a basis for \( V \) if
\[ \text{span} (v_1, \ldots, v_n) = V \]
and \( \{v_1, \ldots, v_n\} \) is linearly independent. The set of vectors is linearly dependent if it is not linearly independent.

The next theorem is called the exchange theorem. It is very important that you understand this theorem. It is so important that I have given three proofs of it. The first two proofs amount to the same thing but are worded slightly differently.

Theorem C.1.4 Let \( \{x_1, \ldots, x_r\} \) be a linearly independent set of vectors such that each \( x_i \) is in the span\( \{y_1, \ldots, y_s\} \). Then \( r \leq s \).

Proof: Define \( \text{span} \{y_1, \ldots, y_s\} = V \), it follows there exist scalars, \( c_1, \ldots, c_s \) such that
\[ x_1 = \sum_{i=1}^{s} c_i y_i. \] (3.5)

Not all of these scalars can equal zero because if this were the case, it would follow that \( x_1 = 0 \) and so \( \{x_1, \ldots, x_r\} \) would not be linearly independent. Indeed, if \( x_1 = 0, 1x_1 + \sum_{i=2}^{r} 0x_i = x_1 = 0 \) and so there would exist a nontrivial linear combination of the vectors \( \{x_1, \ldots, x_r\} \) which equals zero.

Say \( c_k \neq 0 \). Then solve ((3.5)) for \( y_k \) and obtain \( y_k \in \text{span} \left( x_1, y_1, \ldots, y_{k-1}, y_{k+1}, \ldots, y_s \right) \).

Define \( \{z_1, \ldots, z_{s-1}\} \) by
\[ \{z_1, \ldots, z_{s-1}\} \equiv \{y_1, \ldots, y_{k-1}, y_{k+1}, \ldots, y_s\} \]

Therefore, \( \text{span} \{x_1, z_1, \ldots, z_{s-1}\} = V \) because if \( v \in V \), there exist constants \( c_1, \ldots, c_s \) such that
\[ v = \sum_{i=1}^{s-1} c_i z_i + c_s y_k. \]

Now replace the \( y_k \) in the above with a linear combination of the vectors, \( \{x_1, z_1, \ldots, z_{s-1}\} \) to obtain \( v \in \text{span} \{x_1, z_1, \ldots, z_{s-1}\} \). The vector \( y_k \), in the list \( \{y_1, \ldots, y_s\} \), has now been replaced with the vector \( x_1 \) and the resulting modified list of vectors has the same span as the original list of vectors, \( \{y_1, \ldots, y_s\} \).

Now suppose that \( r > s \) and that \( \text{span} \{x_1, \ldots, x_r, z_1, \ldots, z_p\} = V \) where the vectors, \( z_1, \ldots, z_p \) are each taken from the set, \( \{y_1, \ldots, y_s\} \) and \( l + p = s \). This has now been done for \( l = 1 \) above. Then since \( r > s \), it follows that \( l \leq s < r \) and so \( l + 1 \leq r \). Therefore, \( x_{l+1} \) is a vector not in the list, \( \{x_1, \ldots, x_l\} \) and since \( \text{span} \{x_1, \ldots, x_l, z_1, \ldots, z_p\} = V \) there exist scalars, \( c_i \) and \( d_j \) such that
\[ x_{l+1} = \sum_{i=1}^{l} c_i x_i + \sum_{j=1}^{p} d_j z_j. \] (3.6)
C.1. VECTOR SPACES

Now not all the \( d_j \) can equal zero because if this were so, it would follow that \( \{x_1, \ldots, x_r\} \)
would be a linearly dependent set because one of the vectors would equal a linear combination of the others. Therefore, \((3.6)\) can be solved for one of the \( z_i \), say \( z_k \), in terms of \( x_{i+1} \) and the other \( z_i \) and just as in the above argument, replace that \( z_i \) with \( x_{i+1} \) to obtain

\[
\text{span} \left( x_1, \ldots, x_{i+1}, z_1, \ldots, z_k-1, z_k+1, \ldots, z_p \right) = V.
\]

Continue this way, eventually obtaining

\[
\text{span} \left( x_1, \ldots, x_r \right) = V.
\]

But then \( x_r \in \text{span} \left( x_1, \ldots, x_s \right) \) contrary to the assumption that \( \{x_1, \ldots, x_r\} \) is linearly independent. Therefore, \( r \leq s \) as claimed.

**Theorem C.1.5** If

\[
\text{span} \left( u_1, \ldots, u_r \right) \subseteq \text{span} \left( v_1, \ldots, v_s \right)
\]

and \( \{u_1, \ldots, u_r\} \) are linearly independent, then \( r \leq s \).

**Proof:** Let \( V \equiv \text{span} \left( v_1, \ldots, v_s \right) \) and suppose \( r > s \). Let \( A_l \equiv \{u_1, \ldots, u_l\}, A_0 = \emptyset \), and let \( B_{s-l} \) denote a subset of the vectors, \( \{v_1, \ldots, v_s\} \) which contains \( s-l \) vectors and has the property that \( \text{span} \left( A_l, B_{s-l} \right) = V \). Note that the assumption of the theorem says \( \text{span} \left( A_0, B_s \right) = V \).

Now an exchange operation is given for \( \text{span} \left( A_l, B_{s-l} \right) = V \). Since \( r > s \), it follows \( l < r \). Letting

\[
B_{s-l} \equiv \{z_1, \ldots, z_{s-l}\} \subseteq \{v_1, \ldots, v_s\},
\]

it follows there exist constants, \( c_i \) and \( d_i \) such that

\[
u_{i+1} = \sum_{i=1}^{l} c_i u_i + \sum_{i=1}^{s-l} d_i z_i,
\]

and not all the \( d_i \) can equal zero. (If they were all equal to zero, it would follow that the set, \( \{u_1, \ldots, u_r\} \) would be dependent since one of the vectors in it would be a linear combination of the others.)

Let \( d_k \neq 0 \). Then \( z_k \) can be solved for as follows.

\[
z_k = \frac{1}{d_k} u_{i+1} - \sum_{i=1}^{l} \frac{c_i}{d_k} u_i - \sum_{i \neq k} \frac{d_i}{d_k} z_i.
\]

This implies \( V = \text{span} \left( A_{l+1}, B_{s-l-1} \right) \), where \( B_{s-l-1} \equiv B_{s-l} \setminus \{z_k\} \), a set obtained by deleting \( z_k \) from \( B_{s-l} \). You see, the process exchanged a vector in \( B_{s-l} \) with one from \( \{u_1, \ldots, u_r\} \) and kept the span the same. Starting with \( V = \text{span} \left( A_0, B_s \right) \), do the exchange operation until \( V = \text{span} \left( A_{s-1}, z \right) \) where \( z \in \{v_1, \ldots, v_s\} \). Then one more application of the exchange operation yields \( V = \text{span} \left( A_s \right) \). But this implies \( u_r \in \text{span} \left( A_s \right) = \text{span} \left( u_1, \ldots, u_s \right) \), contradicting the linear independence of \( \{u_1, \ldots, u_r\} \). It follows that \( r \leq s \) as claimed.

Here is yet another proof in case you didn’t like either of the last two.

**Theorem C.1.6** If

\[
\text{span} \left( u_1, \ldots, u_r \right) \subseteq \text{span} \left( v_1, \ldots, v_s \right)
\]

and \( \{u_1, \ldots, u_r\} \) are linearly independent, then \( r \leq s \).
**Proof:** Suppose $r > s$. Since each $u_k \in \text{span} \{ v_1, \ldots, v_s \}$, it follows $\text{span} \{ u_1, \ldots, u_s, v_1, \ldots, v_s \} = \text{span} \{ v_1, \ldots, v_s \}$. Let $\{ v_{k_1}, \ldots, v_{k_j} \}$ denote a subset of the set $\{ v_1, \ldots, v_s \}$ which has $j$ elements in it. If $j = 0$, this means no vectors from $\{ v_1, \ldots, v_s \}$ are included.

**Claim:** $j = 0$.

**Proof of claim:** Let $j$ be the smallest nonnegative integer such that
\[
\text{span} \{ u_1, \ldots, u_s, v_{k_1}, \ldots, v_{k_j} \} = \text{span} \{ v_1, \ldots, v_s \} \tag{3.7}
\]
and suppose $j \geq 1$. Then since $s < r$, there exist scalars, $a_k$ and $b_i$ such that
\[
u_{s+1} = \sum_{k=1}^{s} a_k u_k + \sum_{i=1}^{j} b_i v_{k_i}.
\]
By linear independence of the $u_k$, not all the $b_i$ can equal zero. Therefore, one of the $v_{k_i}$ is in the span of the other vectors in the above sum. Thus there exist $l_1, \ldots, l_{j-1}$ such that
\[v_{k_i} \in \text{span} \{ u_1, \ldots, u_s, u_{s+1}, v_{l_1}, \ldots, v_{l_{j-1}} \}
\]
and so from (3.7),
\[\text{span} \{ u_1, \ldots, u_s, u_{s+1}, v_{l_1}, \ldots, v_{l_{j-1}} \} = \text{span} \{ v_1, \ldots, v_s \}
\]
contrary to the definition of $j$. Therefore, $j = 0$ and this proves the claim.

It follows from the claim that $\text{span} \{ u_1, \ldots, u_s \} = \text{span} \{ v_1, \ldots, v_s \}$ which implies
\[u_{s+1} \in \text{span} \{ u_1, \ldots, u_s \}
\]
contrary to the assumption the $u_k$ are linearly independent. Therefore, $r \leq s$ as claimed.

**Corollary C.1.7** If $\{ u_1, \ldots, u_m \}$ and $\{ v_1, \ldots, v_n \}$ are two bases for $V$, then $m = n$.

**Proof:** By Theorem C.1.5, $m \leq n$ and $n \leq m$.

**Definition C.1.8** A vector space $V$ is of dimension $n$ if it has a basis consisting of $n$ vectors. This is well defined thanks to Corollary C.1.7. It is always assumed here that $n < \infty$ in this case, such a vector space is said to be finite dimensional.

**Theorem C.1.9** If $V = \text{span} \{ u_1, \ldots, u_n \}$ then some subset of $\{ u_1, \ldots, u_n \}$ is a basis for $V$. Also, if $\{ u_1, \ldots, u_k \} \subseteq V$ is linearly independent and the vector space is finite dimensional, then the set, $\{ u_1, \ldots, u_k \}$, can be enlarged to obtain a basis of $V$.

**Proof:** Let
\[S = \{ E \subseteq \{ u_1, \ldots, u_n \} \text{ such that } \text{span}(E) = V \}.
\]
For $E \in S$, let $|E|$ denote the number of elements of $E$. Let
\[m = \min \{|E| : E \in S\}.
\]
Thus there exist vectors
\[\{ v_1, \ldots, v_m \} \subseteq \{ u_1, \ldots, u_n \}
\]
such that
\[\text{span} \{ v_1, \ldots, v_m \} = V
\]
and \( m \) is as small as possible for this to happen. If this set is linearly independent, it follows it is a basis for \( V \) and the theorem is proved. On the other hand, if the set is not linearly independent, then there exist scalars,

\[
c_1, \ldots, c_m
\]
such that

\[
0 = \sum_{i=1}^{m} c_i v_i
\]

and not all the \( c_i \) are equal to zero. Suppose \( c_k \neq 0 \). Then the vector, \( v_k \), may be solved for in terms of the other vectors. Consequently,

\[
V = \text{span} (v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_m)
\]

contradicting the definition of \( m \). This proves the first part of the theorem.

To obtain the second part, begin with \( \{u_1, \ldots, u_k\} \) and suppose a basis for \( V \) is \( \{v_1, \ldots, v_n\} \). If

\[
\text{span} (u_1, \ldots, u_k) = V,
\]

then \( k = n \). If not, there exists a vector,

\[
u_{k+1} \notin \text{span} (u_1, \ldots, u_k).
\]

Then \( \{u_1, \ldots, u_k, u_{k+1}\} \) is also linearly independent. Continue adding vectors in this way until \( n \) linearly independent vectors have been obtained. Then \( \text{span} (u_1, \ldots, u_n) = V \) because if it did not do so, there would exist \( u_{n+1} \) as just described and \( \{u_1, \ldots, u_n, u_{n+1}\} \) would be a linearly independent set of vectors having \( n + 1 \) elements even though \( \{v_1, \ldots, v_n\} \) is a basis. This would contradict Theorems C.1.5 and C.1.4. Therefore, this list is a basis and this proves the theorem.

It is useful to emphasize some of the ideas used in the above proof.

**Lemma C.1.10** Suppose \( v \notin \text{span} (u_1, \ldots, u_k) \) and \( \{u_1, \ldots, u_k, v\} \) is linearly independent. Then \( \{u_1, \ldots, u_k, v\} \) is also linearly independent.

**Proof:** Suppose \( \sum_{i=1}^{k} c_i u_i + dv = 0 \). It is required to verify that each \( c_i = 0 \) and that \( d = 0 \). But if \( d \neq 0 \), then you can solve for \( v \) as a linear combination of the vectors, \( \{u_1, \ldots, u_k\} \),

\[
v = - \sum_{i=1}^{k} \left( \frac{c_i}{d} \right) u_i
\]

counter to assumption. Therefore, \( d = 0 \). But then \( \sum_{i=1}^{k} c_i u_i = 0 \) and the linear independence of \( \{u_1, \ldots, u_k\} \) implies each \( c_i = 0 \) also. This proves the lemma.

**Theorem C.1.11** Let \( V \) be a nonzero subspace of a finite dimensional vector space, \( W \) of dimension, \( n \). Then \( V \) has a basis with no more than \( n \) vectors.

**Proof:** Let \( v_1 \in V \) where \( v_1 \neq 0 \). If \( \text{span} \{v_1\} = V \), stop. \( \{v_1\} \) is a basis for \( V \). Otherwise, there exists \( v_2 \in V \) which is not in \( \text{span} \{v_1\} \). By Lemma C.1.10 \( \{v_1, v_2\} \) is a linearly independent set of vectors. If \( \text{span} \{v_1, v_2\} = V \), stop. \( \{v_1, v_2\} \) is a basis for \( V \). If \( \text{span} \{v_1, v_2\} \neq V \), then there exists \( v_3 \notin \text{span} \{v_1, v_2\} \) and \( \{v_1, v_2, v_3\} \) is a larger linearly independent set of vectors. Continuing this way, the process must stop before \( n + 1 \) steps because if not, it would be possible to obtain \( n + 1 \) linearly independent vectors contrary to the exchange theorem, Theorems C.1.4 and C.1.5. This proves the theorem.
C.2 Matrix Multiplication As A Linear Transformation

**Definition C.2.1** Let $V$ and $W$ be two finite dimensional vector spaces. A function, $L$ which maps $V$ to $W$ is called a linear transformation and $L \in \mathcal{L}(V,W)$ if for all scalars $\alpha$ and $\beta$, and vectors $v, w$,

$$L(\alpha v + \beta w) = \alpha L(v) + \beta L(w).$$

An example of a linear transformation is familiar matrix multiplication. Let $A = (a_{ij})$ be an $m \times n$ matrix. Then an example of a linear transformation $L : \mathbb{F}^n \to \mathbb{F}^m$ is given by

$$(Lv)_i \equiv \sum_{j=1}^{n} a_{ij} v_j.$$ 

Here

$$v \equiv \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{F}^n.$$

C.3 $\mathcal{L}(V,W)$ As A Vector Space

**Definition C.3.1** 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. What about the dimension of $\mathcal{L}(V,W)$?

**Theorem C.3.2** Let $V$ and $W$ be finite dimensional normed linear spaces of dimension $n$ and $m$ respectively Then $\dim(\mathcal{L}(V,W)) = mn$.

**Proof:** Let the two sets of bases be

$$\{v_1, \cdots, v_n\} \text{ and } \{w_1, \cdots, w_m\}$$

for $X$ and $Y$ respectively. Let $E_{ik} \in \mathcal{L}(V,W)$ be the linear transformation defined on the basis, $\{v_1, \cdots, v_n\}$, by

$$E_{ik} v_j \equiv w_i \delta_{jk}$$

where $\delta_{ik} = 1$ if $i = k$ and 0 if $i \neq k$. Thus

$$E_{ik} v_j \equiv \begin{cases} w_i & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}.$$

Then let $L \in \mathcal{L}(V,W)$. Since $\{w_1, \cdots, w_m\}$ is a basis, there exist constants, $d_{jk}$ such that

$$Lv_i = \sum_{j=1}^{m} d_{ij} w_j.$$
Also

\[ \sum_{j=1}^{m} \sum_{k=1}^{n} d_{jk} E_{jk} (v_r) = \sum_{j=1}^{m} d_{jr} w_j. \]

It follows that

\[ L = \sum_{j=1}^{m} \sum_{k=1}^{n} d_{jk} E_{jk} \]

because the two linear transformations agree on a basis. Since \( L \) is arbitrary this shows

\[ \{ E_{ik} : i = 1, \ldots, m, \ k = 1, \ldots, n \} \]

spans \( L(V, W) \).

If

\[ \sum_{i,k} d_{ik} E_{ik} = 0, \]

then

\[ 0 = \sum_{i,k} d_{ik} E_{ik} (v_l) = \sum_{i=1}^{m} d_{il} w_i \]

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 the dimension of \( L(V, W) \) is \( mn \) as claimed.

\section{Eigenvalues And Eigenvectors Of Linear Transformations}

Let \( V \) be a finite dimensional vector space. For example, it could be a subspace of \( \mathbb{C}^n \). Also suppose \( A \in L(V, V) \). Does \( A \) have eigenvalues and eigenvectors just like the case where \( A \) is a \( n \times n \) matrix?

\begin{theorem}
Let \( V \) be a nonzero finite dimensional complex vector space of dimension \( n \). Suppose also the field of scalars equals \( \mathbb{C} \). Suppose \( A \in L(V, V) \). Then there exists \( v \neq 0 \) and \( \lambda \in \mathbb{C} \) such that

\[ Av = \lambda v. \]

\end{theorem}

Prove: Consider the linear transformations, \( I, A, A^2, \ldots, A^{n^2} \). There are \( n^2 + 1 \) of these transformations and so by Theorem C.3.2 the set is linearly dependent. Thus there exist constants, \( c_i \in \mathbb{C} \) such that

\[ c_0 I + \sum_{k=1}^{n^2} c_k A^k = 0. \]

This implies there exists a polynomial, \( q(\lambda) \) which has the property that \( q(A) = 0 \). In fact, \( q(\lambda) \equiv c_0 + \sum_{k=1}^{n^2} c_k \lambda^k \). Dividing by the leading term, it can be assumed this polynomial is of the form \( \lambda^m + c_{m-1} \lambda^{m-1} + \cdots + c_1 \lambda + c_0 \), a monic polynomial. Now consider all such monic polynomials, \( q \) such that \( q(A) = 0 \) and pick one which has the smallest degree. This

\footnote{All that is really needed is that the minimal polynomial can be completely factored in the given field. The complex numbers have this property from the fundamental theorem of algebra.}
is called the minimal polynomial and will be denoted here by \( p(\lambda) \). By the fundamental theorem of algebra, \( p(\lambda) \) is of the form

\[
p(\lambda) = \prod_{k=1}^{p} (\lambda - \lambda_k).
\]

Thus, since \( p \) has minimal degree,

\[
\prod_{k=1}^{p} (A - \lambda_k I) = 0, \quad \text{but} \quad \prod_{k=1}^{p-1} (A - \lambda_k I) \neq 0.
\]

Therefore, there exists \( u \neq 0 \) such that

\[
v \equiv \left( \prod_{k=1}^{p-1} (A - \lambda_k I) \right) (u) \neq 0.
\]

But then

\[
(A - \lambda_p I) v = (A - \lambda_p I) \left( \prod_{k=1}^{p-1} (A - \lambda_k I) \right) (u) = 0.
\]

This proves the theorem.

**Corollary C.4.2** In the above theorem, each of the scalars, \( \lambda_k \) has the property that there exists a nonzero \( v \) such that \( (A - \lambda_k I) v = 0 \). Furthermore the \( \lambda_i \) are the only scalars with this property.

**Proof:** For the first claim, just factor out \( (A - \lambda_k I) \) instead of \( (A - \lambda_p I) \). Next suppose \( (A - \mu I) v = 0 \) for some \( \mu \) and \( v \neq 0 \). Then

\[
0 = \prod_{k=1}^{p} (A - \lambda_k I) v = \prod_{k=1}^{p-1} (A - \lambda_k I) (Av - \lambda_p v)
\]

\[
= (\mu - \lambda_p) \left( \prod_{k=1}^{p-1} (A - \lambda_k I) \right) v
\]

\[
= (\mu - \lambda_p) \left( \prod_{k=1}^{p-2} (A - \lambda_k I) \right) (A - \lambda_{p-1} I) v
\]

\[
= (\mu - \lambda_p) (\mu - \lambda_{p-1}) \left( \prod_{k=1}^{p-2} (A - \lambda_k I) \right)
\]

continuing this way yields

\[
= \prod_{k=1}^{p} (\mu - \lambda_k) v,
\]

a contradiction unless \( \mu = \lambda_k \) for some \( k \).

Therefore, these are eigenvectors and eigenvalues with the usual meaning and the \( \lambda_k \) are all of the eigenvalues.

**Definition C.4.3** For \( A \in \mathcal{L}(V, V) \) where \( \dim(V) = n \), the scalars, \( \lambda_k \) in the minimal polynomial,

\[
p(\lambda) = \prod_{k=1}^{p} (\lambda - \lambda_k)
\]
are called the eigenvalues of \(A\). The collection of eigenvalues of \(A\) is denoted by \(\sigma(A)\). For \(\lambda\) an eigenvalue of \(A \in L(V, V)\), the generalized eigenspace is defined as
\[
V_\lambda = \{ x \in V : (A - \lambda I)^m x = 0 \text{ for some } m \in \mathbb{N} \}
\]
and the eigenspace is defined as
\[
\{ x \in V : (A - \lambda I) x = 0 \} \equiv \ker(A - \lambda I).
\]
Also, for subspaces of \(V\), \(V_1, V_2, \ldots, V_r\), the symbol, \(V_1 + V_2 + \cdots + V_r\) or the shortened version, \(\sum_{i=1}^{r} V_i\) will denote the set of all vectors of the form \(\sum_{i=1}^{r} v_i\) where \(v_i \in V_i\).

**Lemma C.4.4** The generalized eigenspace for \(\lambda \in \sigma(A)\) where \(A \in L(V, V)\) for \(V\) an \(n\) dimensional vector space is a subspace of \(V\) satisfying
\[
A : V_\lambda \to V_\lambda,
\]
and there exists a smallest integer, \(m\) with the property that
\[
\ker (A - \lambda I)^m = \{ x \in V : (A - \lambda I)^m x = 0 \text{ for some } m \in \mathbb{N} \}. \tag{3.8}
\]

**Proof:** The claim that the generalized eigenspace is a subspace is obvious. To establish the second part, note that
\[
\left\{ \ker (A - \lambda I)^k \right\}
\]
yields an increasing sequence of subspaces. Eventually
\[
\dim (\ker (A - \lambda I)^m) = \dim \left( \ker (A - \lambda I)^{m+1} \right)
\]
and so \(\ker (A - \lambda I)^m = \ker (A - \lambda I)^{m+1}\). Now if \(x \in \ker (A - \lambda I)^{m+2}\), then
\[
(A - \lambda I) x \in \ker (A - \lambda I)^{m+1} = \ker (A - \lambda I)^m
\]
and so there exists \(y \in \ker (A - \lambda I)^m\) such that \((A - \lambda I) x = y\) and consequently
\[
(A - \lambda I)^{m+1} x = (A - \lambda I)^m y = 0
\]
showing that \(x \in \ker (A - \lambda I)^{m+1}\). Therefore, continuing this way, it follows that for all \(k \in \mathbb{N}\),
\[
\ker (A - \lambda I)^m = \ker (A - \lambda I)^{m+k}.
\]
Therefore, this shows (3.8).

The following theorem is of major importance and will be the basis for the very important theorems concerning block diagonal matrices.

**Theorem C.4.5** Let \(V\) be a complex vector space of dimension \(n\) and suppose \(\sigma(A) = \{\lambda_1, \ldots, \lambda_k\}\) where the \(\lambda_i\) are the distinct eigenvalues of \(A\). Denote by \(V_i\) the generalized eigenspace for \(\lambda_i\) and let \(r_i\) be the multiplicity of \(\lambda_i\). By this is meant
\[
V_i = \ker (A - \lambda_i I)^{r_i} \tag{3.9}
\]
and \(r_i\) is the smallest integer with this property. Then
\[
V = \sum_{i=1}^{k} V_i \tag{3.10}
\]
**Proof:** This is proved by induction on $k$. First suppose there is only one eigenvalue, $\lambda_i$ of multiplicity $m$. Then by the definition of eigenvalues given in Definition C.4.3, $A$ satisfies an equation of the form

$$(A - \lambda_1 I)^r = 0$$

where $r$ is as small as possible for this to take place. Thus $\ker(A - \lambda_1 I)^r = V$ and the theorem is proved in the case of one eigenvalue.

Now suppose the theorem is true for any $i \leq k - 1$ where $k \geq 2$ and suppose $\sigma(A) = \{\lambda_1, \ldots, \lambda_k\}$.

**Claim 1:** Let $\mu \neq \lambda_i$. Then $(A - \mu I)^m : V_i \rightarrow V_i$ and is one to one and onto for every $m \in \mathbb{N}$.

**Proof:** It is clear that $(A - \mu I)^m$ maps $V_i$ to $V_i$ because if $v \in V_i$ then $(A - \lambda_i I)^k v = 0$ for some $k \in \mathbb{N}$. Consequently,

$$(A - \lambda_i I)^k (A - \mu I)^m v = (A - \mu I)^m (A - \lambda_i I)^k v = (A - \mu I)^m 0 = 0$$

which shows that $(A - \mu I)^m v \in V_i$.

It remains to verify that $(A - \mu I)^m$ is one to one. This will be done by showing that $(A - \mu I)$ is one to one. Let $w \in V_i$ and suppose $(A - \mu I)w = 0$ so that $Aw = \mu w$. Then for some $m \in \mathbb{N}$, $(A - \lambda_i I)^m w = 0$ and so by the binomial theorem,

$$(\mu - \lambda_i)^m w = \sum_{l=0}^{m} \binom{m}{l} (-\lambda_i)^{m-l} \mu^l w$$

$$(\sum_{l=0}^{m} \binom{m}{l} (-\lambda_i)^{m-l} A^l w) = (A - \lambda_i I)^m w = 0.$$

Therefore, since $\mu \neq \lambda_i$, it follows $w = 0$ and this verifies $(A - \mu I)$ is one to one. Thus $(A - \mu I)^m$ is also one to one on $V_i$. Letting $\{u_1', \ldots, u_{r_k}'\}$ be a basis for $V_i$, it follows $\{(A - \mu I)^m u_1', \ldots, (A - \mu I)^m u_{r_k}'\}$ is also a basis and so $(A - \mu I)^m$ is also onto.

Let $p$ be the smallest integer such that $\ker(A - \lambda_k I)^p = V_k$ and define

$$W \equiv (A - \lambda_k I)^p (V).$$

**Claim 2:** $A : W \rightarrow W$ and $\lambda_k$ is not an eigenvalue for $A$ restricted to $W$.

**Proof:** Suppose to the contrary that $A (A - \lambda_k I)^p u = \lambda_k (A - \lambda_k I)^p u$ where $(A - \lambda_k I)^p u \neq 0$. Then subtracting $\lambda_k (A - \lambda_k I)^p u$ from both sides yields

$$(A - \lambda_k I)^p+1 u = 0$$

and so $u \in \ker((A - \lambda_k I)^p)$ from the definition of $p$. But this requires $(A - \lambda_k I)^p u = 0$ contrary to $(A - \lambda_k I)^p u \neq 0$. This has verified the claim.

It follows from this claim that the eigenvalues of $A$ restricted to $W$ are a subset of $\{\lambda_1, \ldots, \lambda_{k-1}\}$. Letting

$$V'_i \equiv \left\{w \in W : (A - \lambda_i)^l w = 0 \text{ for some } l \in \mathbb{N}\right\},$$

it follows from the induction hypothesis that

$$W = \sum_{i=1}^{k-1} V'_i \subseteq \sum_{i=1}^{k-1} V_i.$$
C.4. EIGENVALUES AND EIGENVECTORS OF LINEAR TRANSFORMATIONS

From Claim 1, \((A - \lambda_k I)^p\) maps \(V_i\) one to one and onto \(V_i\). Therefore, if \(x \in V_i\), then 
\[
(A - \lambda_k I)^p x \in V_i \quad \text{for each } i.
\]

Consequently
\[
(A - \lambda_k I)^p \left( x - \sum_{i=1}^{k-1} x_i \right) = 0
\]

and so there exists \(x_k \in V_k\) such that
\[
x = \sum_{i=1}^{k-1} x_i + x_k.
\]

This proves the theorem.

**Definition C.4.6**

Let \(\{V_i\}_{i=1}^r\) be subspaces of \(V\) which have the property that if \(v_i \in V_i\) and
\[
\sum_{i=1}^r v_i = 0,
\]

then \(v_i = 0\) for each \(i\). Under this condition, 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.

**Theorem C.4.7**

Let \(\{V_i\}_{i=1}^m\) be subspaces of \(V\) which have the property (3.11) and let \(B_i = \{u_{i,1}, \cdots, u_{i,r_i}\}\) be a basis for \(V_i\). Then \(\{B_1, \cdots, B_m\}\) is a basis for \(V_1 \oplus \cdots \oplus V_m = \sum_{i=1}^m V_i\).

**Proof:** It is clear that \(\text{span}(B_1, \cdots, B_m) = V_1 \oplus \cdots \oplus V_m\). It only remains to verify that \(\{B_1, \cdots, B_m\}\) is linearly independent. Arbitrary elements of \(\text{span}(B_1, \cdots, B_m)\) are of the form
\[
\sum_{k=1}^m \sum_{i=1}^{r_i} c_i^k u_i^k.
\]

Suppose then that
\[
\sum_{k=1}^m \sum_{i=1}^{r_i} c_i^k u_i^k = 0.
\]

Since \(\sum_{i=1}^{r_i} c_i^k u_i^k \in V_k\), it follows \(\sum_{i=1}^{r_i} c_i^k u_i^k = 0\) for each \(k\). But then \(c_i^k = 0\) for each \(i = 1, \cdots, r_i\). This proves the theorem.

The following corollary is the main result.

**Corollary C.4.8**

Let \(V\) be a complex vector space of dimension, \(n\) and let \(A \in \mathcal{L}(V, V)\). Also suppose \(\sigma(A) = \{\lambda_1, \cdots, \lambda_s\}\) where the \(\lambda_i\) are distinct. Then letting \(V_{\lambda_i}\) denote the generalized eigenspace for \(\lambda_i\),
\[
V = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_s}
\]

and if \(B_i\) is a basis for \(V_{\lambda_i}\), then \(\{B_1, B_2, \cdots, B_s\}\) is a basis for \(V\).
Proof: It is necessary to verify that the $V_{\lambda_i}$ satisfy condition (3.11). Let $V_{\lambda_i} = \ker (A - \lambda_i I)^r$ and suppose $v_i \in V_{\lambda_i}$ and $\sum_{i=1}^{k} v_i = 0$ where $k \leq s$. It is desired to show this implies each $v_i = 0$. It is clearly true if $k = 1$. Suppose then that the condition holds for $k - 1$ and
\[
\sum_{i=1}^{k} v_i = 0
\]
and not all the $v_i = 0$. By Claim 1 in the proof of Theorem 17.2.2, multiplying by $(A - \lambda_k I)^r$ yields
\[
\sum_{i=1}^{k-1} (A - \lambda_k I)^r v_i = \sum_{i=1}^{k-1} v'_i = 0
\]
where $v'_i \in V_{\lambda_k}$. Now by induction, each $v'_i = 0$ and so each $v_i = 0$ for $i \leq k - 1$. Therefore, the sum, $\sum_{i=1}^{k} v_i$ reduces to $v_k$ and so $v_k = 0$ also.

By Theorem 17.2.2, $\sum_{i=1}^{s} V_{\lambda_i} = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_s} = V$ and by Theorem C.4.7 $\{B_1, B_2, \cdots, B_s\}$ is a basis for $V$. This proves the corollary.

C.5 Block Diagonal Matrices

In this section the vector space will be $\mathbb{C}^n$ and the linear transformations will be $n \times n$ matrices.

Definition C.5.1 Let $A$ and $B$ be two $n \times n$ matrices. Then $A$ is similar to $B$, written as $A \sim B$, when there exists an invertible matrix, $S$ such that $A = S^{-1} BS$.

Theorem C.5.2 Let $A$ be an $n \times n$ matrix. Letting $\lambda_1, \lambda_2, \cdots, \lambda_r$ be the distinct eigenvalues of $A$, arranged in any order, there exist square matrices, $P_1, \cdots, P_r$ such that $A$ is similar to the block diagonal matrix,
\[
P = \begin{pmatrix}
P_1 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & P_r
\end{pmatrix}
\]
in which $P_k$ has the single eigenvalue $\lambda_k$. Denoting by $r_k$ the size of $P_k$, it follows that $r_k$ equals the dimension of the generalized eigenspace for $\lambda_k$,
\[
r_k = \dim \{x : (A - \lambda_k I)^m = 0 \text{ for some } m\} \equiv \dim (V_{\lambda_k})
\]
Furthermore, if $S$ is the matrix satisfying $S^{-1}AS = P$, then $S$ is of the form
\[
(\begin{array}{ccc}
B_1 & \cdots & B_r
\end{array})
\]
where $B_k = (u_{k,1}^1 \cdots u_{k,r_k}^k)$ in which the columns, $\{u_{k,1}^1, \cdots, u_{k,r_k}^k\} = D_k$ constitute a basis for $V_{\lambda_k}$.

Proof: By Corollary C.4.8 $\mathbb{C}^n = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_s}$ and a basis for $\mathbb{C}^n$ is $\{D_1, \cdots, D_r\}$ where $D_k$ is a basis for $V_{\lambda_k}$.

Let
\[
S = \begin{pmatrix}
B_1 & \cdots & B_r
\end{pmatrix}
\]
where the $B_i$ are the matrices described in the statement of the theorem. Then $S^{-1}$ must be of the form
\[
S^{-1} = \begin{pmatrix}
C_1 \\
\vdots \\
C_r
\end{pmatrix}
\]
where $C_i B_i = I_{r_i \times r_i}$. Also, if $i \neq j$, then $C_i AB_j = 0$ the last claim holding because $A : V_j \rightarrow V_j$ so the columns of $AB_j$ are linear combinations of the columns of $B_j$ and each of these columns is orthogonal to the rows of $C_i$. Therefore,

$$S^{-1}AS = \begin{pmatrix} C_1 & \vdots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\
C_r & \vdots & \vdots & \vdots \\ \end{pmatrix} \begin{pmatrix} B_1 & \cdots & B_r \end{pmatrix}$$

$$= \begin{pmatrix} C_1 AB_1 & 0 & \cdots & 0 \\
0 & C_2 AB_2 & \cdots & 0 \\
\vdots & 0 & \ddots & \vdots \\
0 & \cdots & 0 & C_r AB_r \end{pmatrix}$$

and $C_r AB_r$ is an $r_k \times r_k$ matrix.

What about the eigenvalues of $C_r AB_r$? The only eigenvalue of $A$ restricted to $V_{\lambda_k}$ is $\lambda_k$ because if $Ax = \mu x$ for some $x \in V_{\lambda_k}$ and $\mu \neq \lambda_k$, then as in Claim 1 of Theorem 17.2.2,

$$(A - \lambda_k I)^{r_k} x \neq 0$$

contrary to the assumption that $x \in V_{\lambda_k}$. Suppose then that $C_r AB_r x = \lambda x$ where $x \neq 0$. Why is $\lambda = \lambda_k$? Let $y = B_r x$ so $y \in V_{\lambda_k}$. Then

$$S^{-1}Ay = S^{-1}AS \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}$$

and so

$$Ay = \lambda S \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix} = \lambda y.$$  

Therefore, $\lambda = \lambda_k$ because, as noted above, $\lambda_k$ is the only eigenvalue of $A$ restricted to $V_{\lambda_k}$. Now letting $P_k = C_r AB_r$, this proves the theorem.

The above theorem contains a result which is of sufficient importance to state as a corollary.

**Corollary C.5.3** Let $A$ be an $n \times n$ matrix and let $D_k$ denote a basis for the generalized eigenspace for $\lambda_k$. Then $\{D_1, \cdots, D_r\}$ is a basis for $\mathbb{C}^n$.

More can be said. Recall Theorem 10.4.3 on Page 193. From this theorem, there exist unitary matrices, $U_k$ such that $U_k^* P_k U_k = T_k$ where $T_k$ is an upper triangular matrix of the form

$$\begin{pmatrix} \lambda_k & \cdots & * \\
\vdots & \ddots & \vdots \\
0 & \cdots & \lambda_k \end{pmatrix} \equiv T_k$$
Now let $U$ be the block diagonal matrix defined by

$$U = \begin{pmatrix} U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_r \end{pmatrix}.$$ 

By Theorem C.5.2 there exists $S$ such that

$$S^{-1}AS = \begin{pmatrix} P_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_r \end{pmatrix}.$$ 

Therefore,

$$U^*SASU = \begin{pmatrix} \quad U_1^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_r^* \end{pmatrix} \begin{pmatrix} P_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_r \end{pmatrix} \begin{pmatrix} \quad U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_r \end{pmatrix} = \begin{pmatrix} \quad \quad \quad \quad \quad T_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \quad \quad \quad \quad \quad T_r \end{pmatrix}.$$ 

This proves most of the following corollary of Theorem C.5.2.

**Corollary C.5.4** Let $A$ be an $n \times n$ matrix. Then $A$ is similar to an upper triangular, block diagonal matrix of the form

$$T = \begin{pmatrix} T_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & T_r \end{pmatrix}$$

where $T_k$ is an upper triangular matrix having only $\lambda_k$ on the main diagonal. The diagonal blocks can be arranged in any order desired. If $T_k$ is an $m_k \times m_k$ matrix, then

$$m_k = \dim \{ x : (A - \lambda_k I)^m = 0 \text{ for some } m \in \mathbb{N} \}.$$ 

Furthermore, $m_k$ is the multiplicity of $\lambda_k$ as a zero of the characteristic polynomial of $A$.

**Proof:** The only thing which remains is the assertion that $m_k$ equals the multiplicity of $\lambda_k$ as a zero of the characteristic polynomial. However, this is clear from the observation that since $T$ is similar to $A$ they have the same characteristic polynomial because

$$\det (A - \lambda I) = \det (S (T - \lambda I) S^{-1}) = \det (S) \det (S^{-1}) \det (T - \lambda I) = \det (SS^{-1}) \det (T - \lambda I) = \det (T - \lambda I)$$

and the observation that since $T$ is upper triangular, the characteristic polynomial of $T$ is of the form

$$\prod_{k=1}^r (\lambda_k - \lambda)^{m_k}.$$
The above corollary has tremendous significance especially if it is pushed even further resulting in the Jordan Canonical form. This form involves still more similarity transformations resulting in an especially revealing and simple form for each of the $T_k$, but the result of the above corollary is sufficient for most applications.

It is significant because it enables one to obtain great understanding of powers of $A$ by using the matrix $T$. From Corollary C.5.4 there exists an $n \times n$ matrix, $S^2$ such that

$$A = S^{-1}TS.$$ 

Therefore, $A^2 = S^{-1}TSS^{-1}TS = S^{-1}T^2S$ and continuing this way, it follows

$$A^k = S^{-1}T^kS.$$ 

where $T$ is given in the above corollary. Consider $T^k$. By block multiplication,

$$T^k = \begin{pmatrix} T^k_1 & & 0 \\ & \ddots & \\ 0 & & T^k_r \end{pmatrix}.$$ 

The matrix, $T_s$ is an $m_s \times m_s$ matrix which is of the form

$$T_s = \begin{pmatrix} \alpha & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha \end{pmatrix}$$ 

which can be written in the form

$$T_s = D + N$$

for $D$ a multiple of the identity and $N$ an upper triangular matrix with zeros down the main diagonal. Therefore, by the Cayley Hamilton theorem, $N^{m_s} = 0$ because the characteristic equation for $N$ is just $\lambda^{m_s} = 0$. Such a transformation is called nilpotent. You can see $N^{m_s} = 0$ directly also, without having to use the Cayley Hamilton theorem. 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 $k \geq m_s$.

$$T^k_s = (D + N)^k = \sum_{j=0}^{m_s} \binom{k}{j} D^{k-j}N^j$$

$$= \sum_{j=0}^{m_s} \binom{k}{j} D^{k-j}N^j,$$ 

the third equation holding because $N^{m_s} = 0$. Thus $T^k_s$ is of the form

$$T^k_s = \begin{pmatrix} \alpha^k & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha^k \end{pmatrix}.$$ 

\textbf{Lemma C.5.5} Suppose $T$ is of the form $T_s$ described above in (3.12) where the constant, $\alpha$, on the main diagonal is less than one in absolute value. Then

$$\lim_{k \to \infty} (T^k)_{ij} = 0.$$ 

\footnote{The $S$ here is written as $S^{-1}$ in the corollary.}
**Proof:** From (3.13), 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 $\left| (N^j)_{pq} \right|$ for $0 \leq j \leq m_s$,

$$\left| (T^k)_{pq} \right| \leq m_s C \left( \frac{k(k-1)\cdots(k-m_s+1)}{m_s!} \right) |\alpha|^{k-m_s},$$

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.

### C.6 The Matrix Of A Linear Transformation

If $V$ is an $n$ dimensional vector space and $\{v_1, \ldots, v_n\}$ is a basis for $V$, there exists a linear map

$$q : \mathbb{F}^n \to V$$

defined as

$$q (a) \equiv \sum_{i=1}^{n} a_i v_i,$$

where

$$a = \sum_{i=1}^{n} a_i \mathbf{e}_i,$$

for $\mathbf{e}_i$ the standard basis vectors for $\mathbb{F}^n$ consisting of

$$\mathbf{e}_i \equiv \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

where the one is in the $i^{th}$ slot. It is clear that $q$ defined in this way, is one to one, onto, and linear. For $v \in V$, $q^{-1}(v)$ is a list of scalars called the components of $v$ with respect to the basis $\{v_1, \ldots, v_n\}$.

**Definition C.6.1** Given a linear transformation $L$, mapping $V$ to $W$, where $\{v_1, \ldots, v_n\}$ is a basis of $V$ and $\{w_1, \ldots, w_m\}$ is a basis for $W$, an $m \times n$ matrix $A = (a_{ij})$ is called the matrix of the transformation $L$ with respect to the given choice of bases for $V$ and $W$, if whenever $v \in V$, then multiplication of the components of $v$ by $(a_{ij})$ yields the components of $Lv$. 
C.6. THE MATRIX OF A LINEAR TRANSFORMATION

The following diagram is descriptive of the definition. Here \( q_V \) and \( q_W \) are the maps defined above with reference to the bases, \( \{ v_1, \ldots, v_n \} \) and \( \{ w_1, \ldots, w_m \} \) respectively.

\[
\begin{array}{ccc}
L & V & \rightarrow & W \\
q_V & q_W & \uparrow \circ \uparrow & F_n \rightarrow F^m \\
\{ v_1, \ldots, v_n \} & \rightarrow & \{ w_1, \ldots, w_m \} \\
\end{array}
\]  

(3.14)

Letting \( b \in F^n \), this requires

\[
\sum_{i,j} a_{ij} b_j w_i = L \sum_j b_j v_j = \sum_j b_j L v_j.
\]

Now

\[
L v_j = \sum_i c_{ij} w_i
\]

for some choice of scalars \( c_{ij} \) because \( \{ w_1, \ldots, w_m \} \) is a basis for \( W \). Hence

\[
\sum_{i,j} a_{ij} b_j w_i = \sum_j b_j \sum_i c_{ij} w_i = \sum_j c_{ij} b_j w_i.
\]

It follows from the linear independence of \( \{ w_1, \ldots, w_m \} \) that

\[
\sum_j a_{ij} b_j = \sum_j c_{ij} b_j
\]

for any choice of \( b \in F^n \) and consequently

\[
a_{ij} = c_{ij}
\]

where \( c_{ij} \) is defined by (3.15). It may help to write (3.15) in the form

\[
\begin{pmatrix}
L v_1 \\
\vdots \\
L v_n
\end{pmatrix}
= \begin{pmatrix}
w_1 \\
\vdots \\
w_m
\end{pmatrix} C
= \begin{pmatrix}
w_1 \\
\vdots \\
w_m
\end{pmatrix} A
\]

(3.16)

where \( C = (c_{ij}) \), \( A = (a_{ij}) \).

**Example C.6.2** Let

\[
V \equiv \{ \text{polynomials of degree 3 or less} \},
\]

\[
W \equiv \{ \text{polynomials of degree 2 or less} \},
\]

and \( L \equiv D \) where \( D \) is the differentiation operator. A basis for \( V \) is \( \{ 1, x, x^2, x^3 \} \) and a basis for \( W \) is \( \{ 1, x, x^2 \} \).

What is the matrix of this linear transformation with respect to this basis? Using (3.16),

\[
\begin{pmatrix}
0 & 1 & 2x & 3x^2
\end{pmatrix}
= \begin{pmatrix}
1 & x & x^2
\end{pmatrix} C.
\]

It follows from this that

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

Now consider the important case where \( V = F^n \), \( W = F^m \), and the basis chosen is the standard basis of vectors \( e_i \) described above. Let \( L \) be a linear transformation from \( F^n \) to
and let $A$ be the matrix of the transformation with respect to these bases. In this case the coordinate maps $q_V$ and $q_W$ are simply the identity map and the requirement that $A$ is the matrix of the transformation amounts to

$$\pi_i(Lb) = \pi_i(AB)$$

where $\pi_i$ denotes the map which takes a vector in $\mathbb{F}^m$ and returns the $i^{th}$ entry in the vector, the $i^{th}$ component of the vector with respect to the standard basis vectors. Thus, if the components of the vector in $\mathbb{F}^n$ with respect to the standard basis are $(b_1, \ldots, b_n)$,

$$b = \left( b_1 \quad \cdots \quad b_n \right)^T = \sum_i b_i e_i,$$

then

$$\pi_i(Lb) \equiv (Lb)_i = \sum_j a_{ij} b_j.$$

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.

$$\begin{array}{ccc}
\mathbb{F}^n & A_2 & \mathbb{F}^m \\
q_2 & \circ & p_2 \\
V & \circ & W \\
q_1 & \circ & p_1 \\
\mathbb{F}^n & A_1 & \mathbb{F}^m
\end{array}$$

In this diagram $q_i$ and $p_i$ are coordinate maps as described above. From the diagram,

$$p_1^{-1} p_2 A_2 q_2^{-1} q_1 = A_1,$$

where $q_2^{-1} q_1$ and $p_1^{-1} p_2$ are one to one, onto, and linear maps.

**Definition C.6.3** In the special case where $V = W$ and only one basis is used for $V = W$, this becomes

$$q_1^{-1} q_2 A_2 q_2^{-1} q_1 = A_1.$$

Letting $S$ be the matrix of the linear transformation $q_2^{-1} q_1$ with respect to the standard basis vectors in $\mathbb{F}^n$,

$$S^{-1} A_2 S = A_1. \quad (3.17)$$

When this occurs, $A_1$ is said to be similar to $A_2$ and $A \rightarrow S^{-1} AS$ is called a similarity transformation.

Here is some terminology.

**Definition C.6.4** 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 C.6.5** $[x]$ 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 which you should do if you have not seen it.

**Theorem C.6.6** 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 C.6.7** 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, $A \sim B$ and pick a basis for $V$,

$$ \{v_1, \ldots, v_n\}. $$

Define $L \in \mathcal{L}(V, V)$ by

$$ L_{v_i} \equiv \sum_j a_{ji}v_j $$

where $A = (a_{ij})$. Then if $B = (b_{ij})$, and $S = (s_{ij})$ is the matrix which provides the similarity transformation,

$$ A = S^{-1}BS, $$

between $A$ and $B$, it follows that

$$ L_{v_i} = \sum_{r,s,j} s_{ir}b_{rs}(s^{-1})_{sj}v_j. \quad (3.18) $$

Now define

$$ w_i \equiv \sum_j (s^{-1})_{ij}v_j. $$

Then from (3.18),

$$ \sum_i (s^{-1})_{ki} L_{v_i} = \sum_{i,j,r,s} (s^{-1})_{ki} s_{ir}b_{rs}(s^{-1})_{sj}v_j $$
and so  
\[ Lw_k = \sum_s b_{ks}w_s. \]

This proves the theorem because the if part of the conclusion was established earlier.

**Definition C.6.8** An \( n \times n \) matrix, \( A \), is diagonalizable if there exists an invertible \( n \times n \) matrix, \( S \) such that \( S^{-1}AS = D \), where \( D \) is a diagonal matrix. Thus \( D \) has zero entries everywhere except on the main diagonal. Write \( \text{diag} (\lambda_1, \cdots, \lambda_n) \) to denote the diagonal matrix having the \( \lambda_i \) down the main diagonal.

The following theorem is of great significance.

**Theorem C.6.9** Let \( A \) be an \( n \times n \) matrix. Then \( A \) is diagonalizable if and only if \( \mathbb{F}^n \) has a basis of eigenvectors of \( A \). In this case, \( S \) of Definition C.6.8 consists of the \( n \times n \) matrix whose columns are the eigenvectors of \( A \) and \( D = \text{diag} (\lambda_1, \cdots, \lambda_n) \).

**Proof:** Suppose first that \( \mathbb{F}^n \) has a basis of eigenvectors, \( \{v_1, \cdots, v_n\} \) where \( Av_i = \lambda_i v_i \).

Then let \( S \) denote the matrix \( (v_1 \cdots v_n) \) and let \( S^{-1} = \begin{pmatrix} u_1^T \\ \vdots \\ u_n^T \end{pmatrix} \) where \( u_i^T v_j = \delta_{ij} \equiv \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \). \( S^{-1} \) exists because \( S \) has rank \( n \). Then from block multiplication,

\[
S^{-1}AS = \begin{pmatrix} u_1^T \\ \vdots \\ u_n^T \end{pmatrix} (Av_1 \cdots Av_n)
\]

\[
= \begin{pmatrix} u_1^T \\ \vdots \\ u_n^T \end{pmatrix} (\lambda_1 v_1 \cdots \lambda_n v_n)
\]

\[
= \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} = D.
\]

Next suppose \( A \) is diagonalizable so \( S^{-1}AS = D \equiv \text{diag} (\lambda_1, \cdots, \lambda_n) \). Then the columns of \( S \) form a basis because \( S^{-1} \) is given to exist. It only remains to verify that these columns of \( A \) are eigenvectors. But letting \( S = (v_1 \cdots v_n) \), \( AS = SD \) and so \( (Av_1 \cdots Av_n) = (\lambda_1 v_1 \cdots \lambda_n v_n) \) which shows that \( Av_i = \lambda_i v_i \). This proves the theorem.

It makes sense to speak of the determinant of a linear transformation as described in the following corollary.

**Corollary C.6.10** Let \( L \in \mathcal{L}(V, V) \) where \( V \) is an \( n \) dimensional vector space and let \( A \) be the matrix of this linear transformation with respect to a basis on \( V \). Then it is possible to define

\[
\det (L) \equiv \det (A).
\]
Proof: Each choice of basis for \( V \) determines a matrix for \( L \) with respect to the basis. If \( A \) and \( B \) are two such matrices, it follows from Theorem C.6.7 that

\[
A = S^{-1}BS
\]

and so

\[
det(A) = det(S^{-1}) det(B) det(S).
\]

But

\[
1 = det(I) = det(S^{-1}S) = det(S^{-1})
\]

and so

\[
det(A) = det(B)
\]

which proves the corollary.

Definition C.6.11 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 \( A(X) \).

The following theorem explains how the rank of \( A \) is related to the rank of the matrix of \( A \).

Theorem C.6.12 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 \( Y \).

Proof: Recall the diagram which describes what is meant by the matrix of \( A \). Here the two bases are as indicated.

\[
\begin{align*}
\{v_1, \ldots, v_n\} & \quad X \quad A \quad Y \quad \{w_1, \ldots, w_m\} \\
q_X \downarrow & \quad \circ \quad \uparrow q_Y \\
F^n & \quad M \quad \overrightarrow{F^n}
\end{align*}
\]

Let \( \{z_1, \ldots, z_r\} \) be a basis for \( A(X) \). Then since the linear transformation, \( q_Y \) is one to one and onto, \( \{q_Y^{-1}z_1, \ldots, q_Y^{-1}z_r\} \) is a linearly independent set of vectors in \( F^m \). Let \( Au_i = z_i \).

Then

\[
Mq_X^{-1}u_i = q_Y^{-1}z_i
\]

and so the dimension of \( M(F^n) \geq r \). Now if \( M(F^n) < r \) then there exists

\[
y \in M(F^n) \setminus \text{span} \{q_Y^{-1}z_1, \ldots, q_Y^{-1}z_r\}.
\]

But then there exists \( x \in F^n \) with \( Mx = y \). Hence

\[
y = Mx = q_Y^{-1}Aq_Xx \in \text{span} \{q_Y^{-1}z_1, \ldots, q_Y^{-1}z_r\}
\]

a contradiction. This proves the theorem.

The following result is a summary of many concepts.

Theorem C.6.13 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. \( \det(L) \neq 0 \)

5. If \( Lv = 0 \) then \( v = 0 \).

**Proof:** Suppose first \( L \) is one to one and let \( \{v_i\}_{i=1}^n \) be a basis. Then if \( \sum_{i=1}^n c_i Lv_i = 0 \) 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 \( y = \sum_{i=1}^n c_i Lv_i = L(\sum_{i=1}^n c_i v_i) \). Thus \( L \) is onto. It has been shown that 2.) implies 3.).

Now suppose 3.). Then the operation consisting of multiplication by the matrix of \( L, M_L \), must be onto. However, the vectors in \( F^n \) so obtained, consist of linear combinations of the columns of \( M_L \). Therefore, the column rank of \( M_L \) is \( n \). By Theorem 4.4.20 this equals the determinant rank and so \( \det(M_L) \equiv \det(L) \neq 0 \).

Now assume 4.) If \( Lv = 0 \) for some \( v \neq 0 \), it follows that \( M_Lx = 0 \) for some \( x \neq 0 \). Therefore, the columns of \( M_L \) are linearly dependent and so by Theorem 4.4.20, \( \det(M_L) = \det(L) = 0 \) contrary to 4.). Therefore, 4.) implies 5.).

Now suppose 5.) and suppose \( Lv = Lw \). Then \( L(v - w) = 0 \) and so by 5.), \( v - w = 0 \) showing that \( L \) is one to one.

### C.7 The Jordan Canonical Form

Recall Corollary C.5.4. For convenience, this corollary is stated below.

**Corollary C.7.1** Let \( A \) be an \( n \times n \) matrix. Then \( A \) is similar to an upper triangular, block diagonal matrix of the form

\[
T \equiv \begin{pmatrix}
T_1 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & T_r
\end{pmatrix}
\]

where \( T_k \) is an upper triangular matrix having only \( \lambda_k \) on the main diagonal. The diagonal blocks can be arranged in any order desired. If \( T_k \) is an \( m_k \times m_k \) matrix, then

\[
m_k = \dim \{ x : (A - \lambda_k I)^m = 0 \text{ for some } m \in \mathbb{N} \}.
\]

The Jordan Canonical form involves a further reduction in which the upper triangular matrices, \( T_k \) assume a particularly revealing and simple form.

**Definition C.7.2** \( J_k(\alpha) \) is a Jordan block if it is a \( k \times k \) matrix of the form

\[
J_k(\alpha) = \begin{pmatrix}
\alpha & 1 & 0 \\
0 & \ddots & \ddots \\
0 & \ddots & 1 \\
0 & \cdots & 0 & \alpha
\end{pmatrix}
\]

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. A matrix is strictly upper triangular if it is of the form

\[
\begin{pmatrix}
0 & * & * \\
\vdots & \ddots & * \\
0 & \cdots & 0
\end{pmatrix}
\]
where there are zeroes on the main diagonal and below the main diagonal.

The Jordan canonical form involves each of the upper triangular matrices in the conclusion of Corollary C.5.4 being a block diagonal matrix with the blocks being Jordan blocks in which the size of the blocks decreases from the upper left to the lower right. The idea is to show that every square matrix is similar to a unique such matrix which is in Jordan canonical form.

Note that in the conclusion of Corollary C.5.4 each of the triangular matrices is of the form \( \alpha I + N \) where \( N \) is a strictly upper triangular matrix. The existence of the Jordan canonical form follows quickly from the following lemma.

**Lemma C.7.3** Let \( N \) be an \( n \times n \) matrix which is strictly upper triangular. Then there exists an invertible matrix, \( S \) such that

\[
S^{-1}NS = \begin{pmatrix}
J_{r_1}(0) & 0 & \cdots & 0 \\
0 & J_{r_2}(0) & \cdots & 0 \\
& \ddots & \ddots & \ddots \\
0 & 0 & \cdots & J_{r_s}(0)
\end{pmatrix}
\]

where \( r_1 \geq r_2 \geq \cdots \geq r_s \geq 1 \) and \( \sum_{i=1}^{s} r_i = n \).

**Proof:** First note the only eigenvalue of \( N \) is 0. Let \( v_1 \) be an eigenvector. Then \( \{v_1, v_2, \cdots, v_r\} \) is called a chain based on \( v_1 \) if \( Nv_{k+1} = v_k \) for all \( k = 1, 2, \cdots, r \). It will be called a maximal chain if there is no solution, \( v \), to the equation, \( Nv = v_r \).

**Claim 1:** The vectors in any chain are linearly independent and for \( \{v_1, v_2, \cdots, v_r\} \) a chain based on \( v_1 \),

\[
N : \text{span} (v_1, v_2, \cdots, v_r) \rightarrow \text{span} (v_1, v_2, \cdots, v_r). \tag{3.19}
\]

Also if \( \{v_1, v_2, \cdots, v_r\} \) is a chain, then \( r \leq n \).

**Proof:** First note that (3.19) is obvious because

\[
N \sum_{i=1}^{r} c_i v_i = \sum_{i=2}^{r} c_i v_{i-1}.
\]

It only remains to verify the vectors of a chain are independent. If this is not true, you could consider the set of all dependent chains and pick one, \( \{v_1, v_2, \cdots, v_r\} \), which is shortest. Thus \( \{v_1, v_2, \cdots, v_r\} \) is a chain which is dependent and \( r \) is as small as possible. Suppose then that \( \sum_{i=1}^{r} c_i v_i = 0 \) and not all the \( c_i = 0 \). It follows from \( r \) being the smallest length of any dependent chain that all the \( c_i \neq 0 \). Now \( 0 = N^{r-1} (\sum_{i=1}^{r} c_i v_i) = c_1 v_1 \) showing that \( c_1 = 0 \), a contradiction. Therefore, the last claim is obvious. This proves the claim.

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, \( \{v_1^1, \cdots, v_{r_1}^1\} = B_1 \). 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, \( B_2 = \{v_1^2, \cdots, v_{r_2}^2\} \) 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, \( B_3 = \{v_1^3, \cdots, v_{r_3}^3\} \) such that \( r_3 \) is as large as possible. Continue this way. Thus \( r_k \geq r_{k+1} \).

**Claim 2:** The above process terminates with a finite list of chains, \( \{B_1, \cdots, B_s\} \) because for any \( k \), \( \{B_1, \cdots, B_k\} \) is linearly independent.

**Proof of Claim 2:** It suffices to verify that \( \{B_1, \cdots, B_k\} \) is linearly independent. This will be accomplished if it can be shown that no vector may be written as a linear combination
of the other vectors. Suppose then that \( j \) is such that \( v^j_i \) is a linear combination of the other vectors in \( \{ B_1, \ldots, B_k \} \) and that \( j \leq k \) is as large as possible for this to happen. Also suppose that of the vectors, \( v^j_i \in B_j \) such that this occurs, \( i \) is as large as possible. Then

\[
v^j_i = \sum_{q=1}^{p} c_q w_q
\]

where the \( w_q \) are vectors of \( \{ B_1, \ldots, B_k \} \) not equal to \( v^j_i \). Since \( j \) is as large as possible, it follows all the \( w_q \) come from \( \{ B_1, \ldots, B_j \} \) and that those vectors, \( v^j_l \) which are from \( B_j \) have the property that \( l < i \).

Therefore,

\[
v^j_1 = N^{i-1} v^j_i = \sum_{q=1}^{p} c_q N^{i-1} w_q
\]

and this last sum consists of vectors in \( \text{span} (B_1, \ldots, B_{j-1}) \) contrary to the above construction. Therefore, this proves the claim.

**Claim 3:** Suppose \( Nw = 0 \). Then there exists scalars, \( c_i \) such that

\[
w = \sum_{i=1}^{s} c_i v^i_1.
\]

Recall that \( v^i_1 \) is the eigenvector in the \( i^{th} \) chain on which this chain is based.

**Proof of Claim 3:** From the construction, \( w \in \text{span} (B_1, \ldots, B_s) \). Therefore,

\[
w = \sum_{i=1}^{s} \sum_{k=1}^{r_i} c^i_k v^i_k.
\]

Now apply \( N \) to both sides to find

\[
0 = \sum_{i=1}^{s} \sum_{k=2}^{r_i} c^i_k v^i_{k-1}
\]

and so \( c^i_k = 0 \) if \( k \geq 2 \). Therefore,

\[
w = \sum_{i=1}^{s} c^i_1 v^i_1
\]

and this proves the claim.

It remains to verify that \( \text{span} (B_1, \ldots, B_s) = \mathbb{F}^n \). Suppose \( w \notin \text{span} (B_1, \ldots, B_s) \). Since \( N^n = 0 \), there exists a smallest integer, \( k \) such that \( N^k w = 0 \) but \( N^{k-1} w \neq 0 \). Then \( k \leq \min (r_1, \ldots, r_s) \) because there exists a chain of length \( k \) based on the eigenvector, \( N^{k-1} w \), namely

\[
N^{k-1} w, N^{k-2} w, N^{k-3} w, \ldots, w
\]

and this chain must be no longer than the preceding chains. Since \( N^{k-1} w \) is an eigenvector, it follows from Claim 3 that

\[
N^{k-1} w = \sum_{i=1}^{s} c_i v^i_1 = \sum_{i=1}^{s} c_i N^{k-1} v^i_k.
\]

Therefore,

\[
N^{k-1} \left( w - \sum_{i=1}^{s} c_i v^i_k \right) = 0
\]
and so,
\[ NN^{k-2} \left( w - \sum_{i=1}^{s} c_i v_i^k \right) = 0 \]
which implies by Claim 3 that
\[ N^{k-2} \left( w - \sum_{i=1}^{s} c_i v_i^k \right) = \sum_{i=1}^{s} d_i v_i^1 = \sum_{i=1}^{s} d_i N^{k-2} v_{k-1}^i \]
and so
\[ N^{k-2} \left( w - \sum_{i=1}^{s} c_i v_i^k \right) = 0. \]
Continuing this way it follows that for each \( j < k \), there exists a vector, \( z_j \in \text{span} \left( B_1, \ldots, B_s \right) \) such that
\[ N^{k-j} (w - z_j) = 0. \]
In particular, taking \( j = (k - 1) \) yields
\[ N (w - z_{k-1}) = 0 \]
and now using Claim 3 again yields \( w \in \text{span} \left( B_1, \ldots, B_s \right) \), a contradiction. Therefore, \( \text{span} \left( B_1, \ldots, B_s \right) = \mathbb{F}^n \) after all and so \( \{B_1, \ldots, B_s\} \) is a basis for \( \mathbb{F}^n \).

Now consider the block matrix,
\[ S = \begin{pmatrix} B_1 & \cdots & B_s \end{pmatrix} \]
where
\[ B_k = \begin{pmatrix} v_1^k & \cdots & v_{r_k}^k \end{pmatrix}. \]
Thus
\[ S^{-1} = \begin{pmatrix} C_1 \\ \vdots \\ C_s \end{pmatrix} \]
where \( C_i B_i = I_{r_i \times r_i} \) and \( C_i N B_j = 0 \) if \( i \neq j \). Let
\[ C_k = \begin{pmatrix} u_1^T \\ \vdots \\ u_{r_k}^T \end{pmatrix}. \]
Then
\[ C_k N B_k = \begin{pmatrix} u_1^T \\ \vdots \\ u_{r_k}^T \end{pmatrix} \begin{pmatrix} N v_1^k & \cdots & N v_{r_k}^k \end{pmatrix} = \begin{pmatrix} u_1^T \\ \vdots \\ u_{r_k}^T \end{pmatrix} \begin{pmatrix} 0 & v_1^k & \cdots & v_{r_k-1}^k \end{pmatrix} \]
which equals an \( r_k \times r_k \) matrix of the form
\[ J_{r_k} (0) = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & 1 & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix} \]
That is, it has ones down the super diagonal and zeros everywhere else. It follows

\[
S^{-1}NS = \begin{pmatrix}
C_1 & & & \\
& \cdots & & \\
C_s & & \end{pmatrix} N \begin{pmatrix} B_1 & \cdots & B_s \end{pmatrix}
= \begin{pmatrix}
J_{r_1}(0) & 0 & \\
& \ddots & \\
0 & & J_{r_s}(0)
\end{pmatrix}
\]

as claimed. This proves the lemma.

Now let the upper triangular matrices, \(T_k\) be given in the conclusion of Corollary C.5.4. Thus, as noted earlier,

\[
T_k = \lambda_k I_{r_k \times r_k} + N_k
\]

where \(N_k\) is a strictly upper triangular matrix of the sort just discussed in Lemma C.7.3. Therefore, there exists \(S_k\) such that \(S_k^{-1}N_kS_k\) is of the form given in Lemma C.7.3. Now \(S_k^{-1}\lambda_k I_{r_k \times r_k}S_k = \lambda_k I_{r_k \times r_k}\) and so \(S_k^{-1}T_kS_k\) is of the form

\[
\begin{pmatrix}
J_{i_1}(\lambda_k) & 0 & \\
& \ddots & \\
0 & & J_{i_s}(\lambda_k)
\end{pmatrix}
\]

where \(i_1 \geq i_2 \geq \cdots \geq i_s\) and \(\sum_{j=1}^s i_j = r_k\). This proves the following corollary.

**Corollary C.7.4** Suppose \(A\) is an upper triangular \(n \times n\) matrix having \(\alpha\) in every position on the main diagonal. Then there exists an invertible matrix, \(S\) such that

\[
S^{-1}AS = \begin{pmatrix}
J_{k_1}(\alpha) & 0 & \\
& \ddots & \\
0 & & J_{k_r}(\alpha)
\end{pmatrix}
\]

where \(k_1 \geq k_2 \geq \cdots \geq k_r \geq 1\) and \(\sum_{i=1}^r k_i = n\).

The next theorem is the one about the existence of the Jordan canonical form.

**Theorem C.7.5** Let \(A\) be an \(n \times n\) matrix having eigenvalues \(\lambda_1, \cdots, \lambda_r\) where the multiplicity of \(\lambda_i\) as a zero of the characteristic polynomial equals \(m_i\). Then there exists an invertible matrix, \(S\) such that

\[
S^{-1}AS = \begin{pmatrix}
J(\lambda_1) & 0 & \\
& \ddots & \\
0 & & J(\lambda_r)
\end{pmatrix}
\]  \hfill (3.20)

where \(J(\lambda_k)\) is an \(m_k \times m_k\) matrix of the form

\[
\begin{pmatrix}
J_{k_1}(\lambda_k) & 0 & \\
& \ddots & \\
0 & & J_{k_r}(\lambda_k)
\end{pmatrix}
\]  \hfill (3.21)

where \(k_1 \geq k_2 \geq \cdots \geq k_r \geq 1\) and \(\sum_{i=1}^r k_i = m_k\).
Proof: From Corollary C.5.4, there exists S such that $S^{-1}AS$ is of the form

$$T = \begin{pmatrix} T_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & T_r \end{pmatrix}$$

where $T_k$ is an upper triangular $m_k \times m_k$ matrix having only $\lambda_k$ on the main diagonal. By Corollary C.7.4 there exist matrices, $S_k$ such that $S_k^{-1}T_kS_k = J(\lambda_k)$ where $J(\lambda_k)$ is described in (3.21). Now let $M$ be the block diagonal matrix given by

$$M = \begin{pmatrix} S_1 & 0 \\ \vdots & \ddots \\ 0 & S_r \end{pmatrix}.$$ 

It follows that $M^{-1}S^{-1}ASM = M^{-1}TM$ and this is of the desired form. This proves the theorem.

What about the uniqueness of the Jordan canonical form? Obviously if you change the order of the eigenvalues, you get a different Jordan canonical form but it turns out that if the order of the eigenvalues is the same, then the Jordan canonical form is unique. In fact, it is the same for any two similar matrices.

Theorem C.7.6 Let $A$ and $B$ be two similar matrices. Let $J_A$ and $J_B$ be Jordan forms of $A$ and $B$ respectively, made up of the blocks $J_A(\lambda_i)$ and $J_B(\lambda_i)$ respectively. Then $J_A$ and $J_B$ are identical except possibly for the order of the $J(\lambda_i)$ where the $\lambda_i$ are defined above.

Proof: First note that for $\lambda_i$ an eigenvalue, the matrices $J_A(\lambda_i)$ and $J_B(\lambda_i)$ are both of size $m_i \times m_i$ because the two matrices $A$ and $B$, being similar, have exactly the same characteristic equation and the size of a block equals the algebraic multiplicity of the eigenvalue as a zero of the characteristic equation. It is only necessary to worry about the number and size of the Jordan blocks making up $J_A(\lambda_i)$ and $J_B(\lambda_i)$. Let the eigenvalues of $A$ and $B$ be $\{\lambda_1, \ldots, \lambda_r\}$. Consider the two sequences of numbers $\{\text{rank } (A - \lambda I)^m\}$ and $\{\text{rank } (B - \lambda I)^m\}$. Since $A$ and $B$ are similar, these two sequences coincide. (Why?) Also, for the same reason, $\{\text{rank } (J_A(\lambda_i) - \lambda_1 I)^m\}$ coincides with $\{\text{rank } (J_B - \lambda_1 I)^m\}$. Now pick $\lambda_k$ an eigenvalue and consider $\{\text{rank } (J_A - \lambda_k I)^m\}$ and $\{\text{rank } (J_B - \lambda_k I)^m\}$. Then

$$J_A - \lambda_k I = \begin{pmatrix} J_A(\lambda_1 - \lambda_k) & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & J_A(\lambda_r - \lambda_k) \end{pmatrix}$$

and a similar formula holds for $J_B - \lambda_k I$. Here

$$J_A(0) = \begin{pmatrix} J_{k_1}(0) & 0 \\ \vdots & \ddots \\ 0 & \cdots & J_{k_r}(0) \end{pmatrix}$$

and

$$J_B(0) = \begin{pmatrix} J_{l_1}(0) & 0 \\ \vdots & \ddots \\ 0 & \cdots & J_{l_p}(0) \end{pmatrix}.$$
and it suffices to verify that \( l_i = k_i \) for all \( i \). As noted above, \( \sum k_i = \sum l_i \). Now from the above formulas,

\[
\text{rank} \left( J_A - \lambda_k I \right)^m = \sum_{i \neq k} m_i + \text{rank} \left( J_A(0)^m \right)
\]

\[
= \sum_{i \neq k} m_i + \text{rank} \left( J_B(0)^m \right)
\]

\[
= \text{rank} \left( J_B - \lambda_k I \right)^m,
\]

which shows \( \text{rank} \left( J_A(0)^m \right) = \text{rank} \left( J_B(0)^m \right) \) for all \( m \). However,

\[
J_B(0)^m = \begin{pmatrix} J_{l_1}(0)^m & 0 \\ J_{l_2}(0)^m & \ddots \\ 0 & \cdots & J_{l_p}(0)^m \end{pmatrix}
\]

with a similar formula holding for \( J_A(0)^m \) and \( \text{rank} \left( J_A(0)^m \right) = \sum_{i=1}^p \text{rank} \left( J_{l_i}(0)^m \right) \), similar for \( \text{rank} \left( J_B(0)^m \right) \). In going from \( m \) to \( m + 1 \),

\[
\text{rank} \left( J_{l_i}(0)^m \right) - 1 = \text{rank} \left( J_{l_i}(0)^{m+1} \right)
\]

until \( m = l_i \) at which time there is no further change. Therefore, \( p = r \) since otherwise, there would exist a discrepancy right away in going from \( m = 1 \) to \( m = 2 \). Now suppose the sequence \( \{l_i\} \) is not equal to the sequence \( \{k_i\} \). Then \( l_{r-b} \neq k_{r-b} \) for some \( b \) a nonnegative integer taken to be a small as possible. Say \( l_{r-b} > k_{r-b} \). Then, letting \( m = k_{r-b} \),

\[
\sum_{i=1}^r \text{rank} \left( J_{l_i}(0)^m \right) = \sum_{i=1}^r \text{rank} \left( J_{k_i}(0)^m \right)
\]

and in going to \( m + 1 \) a discrepancy must occur because the sum on the right will contribute less to the decrease in rank than the sum on the left. This proves the theorem.

\section*{C.8 Convergence}

\subsection*{C.8.1 The Concept Of A Norm}

Various important numerical techniques in linear algebra have to do with convergence of vectors or matrices. You can’t even begin to logically discuss this without some notion of distance. This is what will be discussed here. First here is the definition of what is meant by a norm.

**Definition C.8.1** Norms satisfy

\[
\|x\| \geq 0, \quad \|x\| = 0 \text{ if and only if } x = 0,
\]

\[
\|x + y\| \leq \|x\| + \|y\|,
\]

\[
\|cx\| = |c| \|x\|
\]

whenever \( c \) is a scalar. A set, \( U \) in \( \mathbb{F}^n \) is **open** if for every \( p \in U \), there exists \( \delta > 0 \) such that

\[
B(p, \delta) \equiv \{x : \|x - p\| < \delta\} \subseteq U.
\]

This is often referred to by saying that every point of the set is an **interior point**.
C.8. CONVERGENCE

To begin with here is a fundamental inequality called the Cauchy Schwarz inequality which is stated here in $\mathbb{C}^n$. First here is a simple lemma.

**Lemma C.8.2** If $z \in \mathbb{C}$ there exists $\theta \in \mathbb{C}$ such that $\theta z = |z|$ and $|\theta| = 1$.

**Proof:** Let $\theta = 1$ if $z = 0$ and otherwise, let $\theta = \frac{z}{|z|}$. Recall that for $z = x + iy$, $\overline{z} = x - iy$.

**Definition C.8.3** For $x \in \mathbb{C}^n$,

$$|x| \equiv \left( \sum_{k=1}^{n} |x_k|^2 \right)^{1/2}.$$

**Theorem C.8.4** (Cauchy Schwarz) The following inequality holds for $x_i$ and $y_i \in \mathbb{C}$.

$$\left| \sum_{i=1}^{n} x_i \overline{y}_i \right| \leq \left( \sum_{i=1}^{n} |x_i|^2 \right)^{1/2} \left( \sum_{i=1}^{n} |y_i|^2 \right)^{1/2}. \quad (3.22)$$

**Proof:** Let $\theta \in \mathbb{C}$ such that $|\theta| = 1$ and

$$\theta \sum_{i=1}^{n} x_i \overline{y}_i = \left| \sum_{i=1}^{n} x_i \overline{y}_i \right|.$$

Thus

$$\theta \sum_{i=1}^{n} x_i \overline{y}_i = \sum_{i=1}^{n} x_i \overline{(\theta y_i)} = \left| \sum_{i=1}^{n} x_i \overline{y}_i \right|.$$

Consider $p(t) \equiv \sum_{i=1}^{n} \left( x_i + t \overline{y}_i \right) \left( \overline{x}_i + t \overline{y}_i \right)$ where $t \in \mathbb{R}$.

$$0 \leq p(t) = \sum_{i=1}^{n} |x_i|^2 + 2t \text{Re} \left( \theta \sum_{i=1}^{n} x_i \overline{y}_i \right) + t^2 \sum_{i=1}^{n} |y_i|^2$$

$$= |x|^2 + 2t \left| \sum_{i=1}^{n} x_i \overline{y}_i \right| + t^2 |y|^2.$$

If $|y| = 0$ then (3.22) is obviously true because both sides equal zero. Therefore, assume $|y| \neq 0$ and then $p(t)$ is a polynomial of degree two whose graph opens up. Therefore, it either has no zeroes, two zeros or one repeated zero. If it has two zeros, the above inequality must be violated because in this case the graph must dip below the $x$ axis. Therefore, it either has no zeros or exactly one. From the quadratic formula this happens exactly when

$$4 \left| \sum_{i=1}^{n} x_i \overline{y}_i \right|^2 = 4 |x|^2 |y|^2 \leq 0$$

and so

$$\left| \sum_{i=1}^{n} x_i \overline{y}_i \right| \leq |x| |y|$$

as claimed. This proves the inequality.
Theorem C.8.5  The norm $|\cdot|$ given in Definition C.8.3 really is a norm. Also if $||\cdot||$ is any norm on $\mathbb{F}^n$. Then $||\cdot||$ is equivalent to $|\cdot|$. That is there exist constants, $\delta$ and $\Delta$ such that

$$\delta ||x|| \leq |x| \leq \Delta ||x||. \quad (3.23)$$

Proof: All of the above properties of a norm are obvious except the second, the triangle inequality. To establish this inequality, use the Cauchy Schwarz inequality to write

$$|\mathbf{x} + \mathbf{y}|^2 \equiv \sum_{i=1}^{n} |x_i + y_i|^2 \leq \sum_{i=1}^{n} |x_i|^2 + \sum_{i=1}^{n} |y_i|^2 + 2 \text{Re} \sum_{i=1}^{n} x_i \bar{y}_i$$

$$\leq |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2 \left( \sum_{i=1}^{n} |x_i|^2 \right)^{1/2} \left( \sum_{i=1}^{n} |y_i|^2 \right)^{1/2}$$

$$= |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2 |\mathbf{x}| |\mathbf{y}| = (|\mathbf{x}| + |\mathbf{y}|)^2$$

and this proves the second property above.

It remains to show the equivalence of the two norms. Letting $\{ \mathbf{e}_k \}$ denote the usual basis vectors for $\mathbb{C}^n$, the Cauchy Schwarz inequality implies

$$||\mathbf{x}|| \equiv \left| \left| \sum_{i=1}^{n} x_i \mathbf{e}_i \right| \right| \leq \left| \sum_{i=1}^{n} |x_i| ||\mathbf{e}_i|| \right| \leq |\mathbf{x}| \left( \sum_{i=1}^{n} ||\mathbf{e}_i||^2 \right)^{1/2}$$

$$\equiv \delta^{-1} |\mathbf{x}|. \quad (3.24)$$

This proves the first half of the inequality.

Suppose the second half of the inequality is not valid. Then there exists a sequence $\mathbf{x}^k \in \mathbb{F}^n$ such that

$$|\mathbf{x}^k| > k \left| \left| \mathbf{x}^k \right| \right|, \quad k = 1, 2, \cdots.$$ Then define

$$\mathbf{y}^k = \frac{\mathbf{x}^k}{|\mathbf{x}^k|}.$$ It follows

$$|\mathbf{y}^k| = 1, \quad |\mathbf{y}^k| > k ||\mathbf{y}^k||$$

and the vector

$$\left( y_1^k, \cdots, y_n^k \right)$$

is a unit vector in $\mathbb{F}^n$. By the Heine Borel theorem from calculus, there exists a subsequence, still denoted by $k$ such that

$$\left( y_1^k, \cdots, y_n^k \right) \rightarrow (y_1, \cdots, y_n) = \mathbf{y}$$ a unit vector. It follows from (3.24) and this that for

$$\mathbf{y} = \sum_{i=1}^{n} y_i \mathbf{e}_i,$$

$$0 = \lim_{k \rightarrow \infty} ||\mathbf{y}^k|| = \lim_{k \rightarrow \infty} \left| \left| \sum_{i=1}^{n} y_i^k \mathbf{e}_i \right| \right| = \left| \left| \sum_{i=1}^{n} y_i \mathbf{e}_i \right| \right|$$

but not all the $y_i$ equal zero because $\mathbf{y}$ is a unit vector. This contradicts the linear independence of $\{ \mathbf{e}_1, \cdots, \mathbf{e}_n \}$ and proves the second half of the inequality.
Corollary C.8.6 Any two norms on $\mathbb{F}^n$ are equivalent. That is, if $\|\cdot\|$ and $\|\|\|\|\|$ are two norms on $\mathbb{F}^n$, then there exist positive constants, $\delta$ and $\Delta$, independent of $x \in X$ such that

$$\delta \|x\| \leq \|x\| \leq \Delta \|x\|.$$  

Proof: By Theorem C.8.5, there are positive constants $\delta_1, \Delta_1, \delta_2, \Delta_2$, all independent of $x \in \mathbb{F}^n$ such that

$$\delta_2 \|x\| \leq |x| \leq \Delta_2 \|x\|,$$

$$\delta_1 \|x\| \leq |x| \leq \Delta_1 \|x\|.$$  

Then

$$\delta_2 \|x\| \leq |x| \leq \Delta_1 \|x\| \leq \frac{\Delta_1 \Delta_2}{\delta_1} \|x\|$$

and so

$$\frac{\delta_2}{\Delta_1} \|x\| \leq |x| \leq \frac{\Delta_2}{\delta_1} \|x\|$$

which proves the corollary.

C.8.2 The Operator Norm

Definition C.8.7 Let norms $\|\cdot\|_{\mathbb{F}_n}$ and $\|\cdot\|_{\mathbb{F}_m}$ be given on $\mathbb{F}^n$ and $\mathbb{F}^m$, respectively. Then $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ denotes the space of $m \times n$ matrices. For $A \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$, the operator norm is defined by

$$\|A\| \equiv \sup \{\|Ax\|_{\mathbb{F}_m} : \|x\|_{\mathbb{F}_n} \leq 1\} < \infty.$$  

Theorem C.8.8 Denote by $\|\cdot\|$ the norm on either $\mathbb{F}^n$ or $\mathbb{F}^m$. The set of $m \times n$ matrices with this norm is a complete normed linear space of dimension $nm$ with

$$\|Ax\| \leq \|A\| \|x\|.$$  

Completeness means that every Cauchy sequence converges.

Proof: It is necessary to show the norm defined on $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ really is a norm. Again the first and third properties listed above for norms are obvious. It remains to show the second and verify $\|A\| < \infty$. There exist constants $\delta, \Delta > 0$ such that

$$\delta \|x\| \leq |x| \leq \Delta \|x\|.$$  

Then,

$$\|A + B\| \equiv \sup \{\|(A + B)(x)\| : \|x\| \leq 1\} \leq \sup \{\|Ax\| : \|x\| \leq 1\} + \sup \{\|Bx\| : \|x\| \leq 1\} \equiv \|A\| + \|B\|.$$  

Next consider the claim that $\|A\| < \infty$. This follows from

$$\|A(x)\| = \left\|A \left( \sum_{i=1}^{n} x_i e_i \right) \right\| \leq \sum_{i=1}^{n} |x_i| \|A(e_i)\| \leq |x| \left( \sum_{i=1}^{n} \|A(e_i)\|^2 \right)^{1/2} \leq \Delta \|x\| \left( \sum_{i=1}^{n} \|A(e_i)\|^2 \right)^{1/2} < \infty.$$
Thus \( ||A|| \leq \Delta \left( \sum_{i=1}^{n} ||A(e_i)||^2 \right)^{1/2} \).

It is clear that a basis for \( L(F^n, F^m) \) consists of matrices of the form \( E_{ij} \) where \( E_{ij} \) consists of the \( m \times n \) matrix having all zeros except for a 1 in the \( ij \)th position. In effect, this considers \( L(F^n, F^m) \) as \( F^{nm} \). Think of the \( m \times n \) matrix as a long vector folded up.

If \( x \neq 0 \),

\[
||Ax|| \frac{1}{||x||} = \left| \left| \frac{Ax}{||x||} \right| \right| \leq ||A||
\]  

(3.25)

It only remains to verify completeness. Suppose then that \( \{ A_k \} \) is a Cauchy sequence in \( L(F^n, F^m) \). Then from (3.25) \( \{ A_kx \} \) is a Cauchy sequence for each \( x \in F^n \). This follows because

\[
||A_kx - A_lx|| \leq ||A_k - A_l|| ||x||
\]

which converges to 0 as \( k, l \to \infty \). Therefore, by completeness of \( F^n \), there exists \( Ax \), the name of the thing to which the sequence, \( \{ A_kx \} \), converges such that

\[
\lim_{k \to \infty} A_kx = Ax.
\]

Then \( A \) is linear because

\[
A(ax + by) = \lim_{k \to \infty} A_k(ax + by) = \lim_{k \to \infty} (aA_kx + bA_ky) = a \lim_{k \to \infty} A_kx + b \lim_{k \to \infty} A_ky = aAx + bAy.
\]

By the first part of this argument, \( ||A|| < \infty \) and so \( A \in L(F^n, F^m) \). This proves the theorem.

It turns out that there are many ways of placing a norm on \( L(F^n, F^m) \) and they are all equivalent. This follows because as noted above, you can think of \( L(F^n, F^m) \) as \( F^{nm} \) and it was shown in Corollary C.8.6 that any two norms on this space are equivalent. One popular norm is the following called the **Frobenius norm**. It is not an operator norm but instead is based on the idea of considering the \( m \times n \) matrix as an element of \( F^{nm} \). Recall the trace of an \( n \times n \) matrix, \( (a_{ij}) \) is just \( \sum_j a_{jj} \). In other words, it is just the sum of the entries on the main diagonal.

**Definition C.8.9** Define an inner product on \( L(F^n, F^m) \) as follows.

\[
(A, B) \equiv tr(AB^\ast)
\]

where \( tr \) denotes the **trace**. Thus

\[
tr(A) \equiv \sum_i A_{ii},
\]

the sum of the entries on the main diagonal. Then define \( ||A|| \equiv (A, A)^{1/2} \). It is obvious this is a norm from the argument above in Theorem C.8.5 applied this time to \( F^{nm} \). This follows because

\[
(A, B) \equiv tr(AB^\ast) \equiv \sum_i \sum_j a_{ij} \overline{b_{ij}}
\]

There are many norms which are used on \( C^n \). The most common ones are listed below. By Corollary C.8.6 they are all equivalent. This means that in any convergence question it does not make any difference which of these norms you use.
Definition C.8.10 Let $x \in \mathbb{C}^n$. Then define for $p \geq 1$,

$$||x||_p \equiv \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

$$||x||_1 \equiv \sum_{i=1}^{n} |x_i|,$n

$$||x||_\infty \equiv \max \{ |x_i|, i = 1, \ldots, n \},$$

$$||x||_2 = \left( \sum_{i=1}^{n} |x_i|^2 \right)^{1/2}.$$

The last is the usual norm often referred to as the Euclidean norm.

It has already been shown that the last of the above norms is really a norm. It is easy to verify that $||\cdot||_1$ is a norm and also not hard to see that $||\cdot||_\infty$ is a norm. You should verify this. The norm, $||\cdot||_p$ is more difficult, however. The following inequality is called Holder’s inequality.

Proposition C.8.11 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 C.8.12 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 $A = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$ and $B = \left( \sum_{i=1}^{n} |y_i|^{p'} \right)^{1/p'}$. Then using Lemma C.8.12,

$$\sum_{i=1}^{n} \frac{|x_i| |y_i|}{AB} \leq \sum_{i=1}^{n} \left[ \frac{1}{p} \left( \frac{|x_i|}{A} \right)^p + \frac{1}{p'} \left( \frac{|y_i|}{B} \right)^{p'} \right] = 1$$

and so

$$\sum_{i=1}^{n} |x_i||y_i| \leq AB = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p} \left( \sum_{i=1}^{n} |y_i|^{p'} \right)^{1/p'}.$$

This proves the proposition.
Theorem C.8.13  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 $\frac{p}{p'} = p - 1$. Then using the Holder inequality,

$$||x + y||^p = \sum_{i=1}^{n} |x_i + y_i|^p$$

$$\leq \sum_{i=1}^{n} |x_i + y_i|^{p-1} |x_i| + \sum_{i=1}^{n} |x_i + y_i|^{p-1} |y_i|$$

$$= \sum_{i=1}^{n} |x_i + y_i|^{p/p'} |x_i| + \sum_{i=1}^{n} |x_i + y_i|^{p/p'} |y_i|$$

$$\leq \left( \sum_{i=1}^{n} |x_i + y_i|^p \right)^{1/p'} \left[ \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p} + \left( \sum_{i=1}^{n} |y_i|^p \right)^{1/p} \right]$$

$$= ||x + y||^{p/p'} \left( ||x||_p + ||y||_p \right)$$

so $||x + y|| \leq ||x||_p + ||y||_p$. This proves the theorem.

It only remains to prove Lemma C.8.12.

Proof of the lemma: Let $p' = q$ to save on notation and consider the following picture:

Note equality occurs when $a^p = b^q$.

Now $||A||_p$ is the operator norm of $A$ taken with respect to $||\cdot||_p$.

Theorem C.8.14  The following holds.

$$||A||_p \leq \left( \sum_k \left( \sum_j |A_{jk}|^p \right)^{q/p} \right)^{1/q}$$

Proof: Let $||x||_p \leq 1$ and let $A = (a_1, \ldots, a_n)$ where the $a_k$ are the columns of $A$. Then

$$Ax = \left( \sum_k x_k a_k \right)$$
and so by Holder’s inequality,

\[ ||Ax||_p \equiv \left| \sum_k x_k a_k \right|_p \leq \sum_k |x_k| ||a_k||_p \leq \left( \sum_k |x_k|^p \right)^{1/p} \left( \sum_k ||a_k||_p^q \right)^{1/q} \]

\[ \leq \left( \sum_k \left( \sum_j |A_{jk}|^p \right)^{q/p} \right)^{1/q} \]

and this shows \( ||A||_p \leq \left( \sum_k \left( \sum_j |A_{jk}|^p \right)^{q/p} \right)^{1/q} \) and proves the theorem.

C.9 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 C.9.1** 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 \( A \).

Before beginning this discussion, it is necessary to define what is meant by convergence in \( L (F^n, F^n) \).

**Definition C.9.2** Let \( \{A_k\}_{k=1}^\infty \) be a sequence in \( L (F^n, F^n) \). 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 \( L (F^n, F^n) \). By Corollary C.8.6 and Theorem C.3.2 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 C.9.3** Suppose \( \{A_k\}_{k=1}^\infty \) is a sequence in \( L (F^n, F^n) \). Then if

\[ \sum_{k=1}^\infty ||A_k|| < \infty, \]

It follows that

\[ \sum_{k=1}^\infty A_k \]

exists. In words, absolute convergence implies convergence.
Proof: For \( p \leq m \leq n \),
\[
\left\| \sum_{k=1}^{n} A_k - \sum_{k=1}^{m} A_k \right\| \leq \sum_{k=p}^{\infty} ||A_k||
\]
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 (3.26) are a Cauchy sequence. It follows that these partial sums converge because as discussed earlier \( \mathcal{L}(\mathbb{F}^n, \mathbb{F}^n) \) can be considered as \( \mathbb{F}^{n^2} \).

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.

Lemma C.9.4 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
\[
1 < \limsup_{p \to \infty} ||A^p||^{1/p}, \quad (3.27)
\]
then \( \sum_{k=0}^{\infty} A^k \) fails to converge.

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.

To see the last claim, if (3.27) holds, then from the first part of this lemma, \( ||A^p|| \) 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.

In this section a significant way to estimate \( \rho(A) \) is presented. It is based on the following lemma.

Lemma C.9.5 If \( |\lambda| > \rho(A) \), for \( A \) an \( n \times n \) matrix, then the series,
\[
\frac{1}{\lambda} \sum_{k=0}^{\infty} \frac{A^k}{\lambda^k}
\]
converges.

Proof: Let \( J \) denote the Jordan canonical form of \( A \). Also, let \( ||A|| \equiv \max \{|a_{ij}|, i, j = 1, 2, \ldots, n\} \).
Then for some invertible matrix, \( S, A = S^{-1}JS \). Therefore,
\[
\frac{1}{\lambda} \sum_{k=0}^{p} \frac{A^k}{\lambda^k} = S^{-1} \left( \frac{1}{\lambda} \sum_{k=0}^{p} \frac{J^k}{\lambda^k} \right) S.
\]
Now from the structure of the Jordan form, \( J = D + N \) where \( D \) is the diagonal matrix consisting of the eigenvalues of \( A \) listed according to algebraic multiplicity and \( N \) is a nilpotent matrix which commutes with \( D \). Say \( N^m = 0 \). Therefore, for \( k \) much larger than \( m \), say \( k > 2m \),

\[
J^k = (D + N)^k = \sum_{l=0}^{m} \binom{k}{l} D^{k-l} N^l.
\]

It follows that

\[
||J^k|| \leq C (m, N) k (k-1) \cdots (k-m+1) ||D||^k
\]

and so

\[
\limsup_{k \to \infty} \left( \frac{||J^k||}{\lambda^k} \right)^{1/k} \leq \lim_{k \to \infty} \left( \frac{C (m, N) k (k-1) \cdots (k-m+1) ||D||^k}{|\lambda|^k} \right)^{1/k} = \frac{||D||}{|\lambda|} < 1.
\]

Therefore, this shows by the root test that \( \sum_{k=0}^{\infty} \left( \frac{||J^k||}{\lambda^k} \right) \) converges. Therefore, by Lemma C.9.3 it follows that

\[
\lim_{k \to \infty} \frac{1}{\lambda} \sum_{i=0}^{k} J^i
\]

exists. In particular this limit exists in every operator norm placed on \( \mathcal{L}(\mathbb{F}^n, \mathbb{F}^n) \), and in particular for every operator norm. Now for any operator norm, \( ||AB|| \leq ||A|| ||B|| \). Therefore,

\[
\left| \mathcal{S}^{-1} \left( \frac{1}{\lambda} \sum_{k=0}^{p} \frac{J^k}{\lambda^k} - \frac{1}{\lambda} \sum_{k=0}^{\infty} \frac{J^k}{\lambda^k} \right) S \right| \leq \left| \mathcal{S}^{-1} \right| ||S|| \left| \left( \frac{1}{\lambda} \sum_{k=0}^{p} \frac{J^k}{\lambda^k} - \frac{1}{\lambda} \sum_{k=0}^{\infty} \frac{J^k}{\lambda^k} \right) \right|
\]

and this converges to 0 as \( p \to \infty \). Therefore,

\[
\frac{1}{\lambda} \sum_{k=0}^{p} \frac{A^k}{\lambda^k} \to \mathcal{S}^{-1} \left( \frac{1}{\lambda} \sum_{k=0}^{\infty} \frac{J^k}{\lambda^k} \right) S
\]

and this proves the lemma.

Actually this lemma is usually accomplished using the theory of functions of a complex variable but the theory involving the Laurent series is not assumed here. In infinite dimensional spaces you have to use complex variable techniques however.

**Lemma C.9.6** Let \( A \) be an \( n \times n \) matrix. Then for any \( ||\cdot|| \), \( \rho (A) \geq \limsup_{p \to \infty} ||A^p||^{1/p} \).

**Proof:** By Lemma C.9.5 and Lemma C.9.4, if \( |\lambda| > \rho (A) \),

\[
\limsup \left( \frac{||A^k||}{|\lambda|^k} \right)^{1/k} \leq 1,
\]

and it doesn’t matter which norm is used because they are all equivalent. Therefore, \( \limsup_{k \to \infty} ||A^k||^{1/k} \leq |\lambda| \). Therefore, since this holds for all \( |\lambda| > \rho (A) \), this proves the lemma.

Now denote by \( \sigma (A)^P \) the collection of all numbers of the form \( \lambda^P \) where \( \lambda \in \sigma (A) \).

**Lemma C.9.7** \( \sigma (A^P) = \sigma (A)^P \)
**Proof:** In dealing with $\sigma(A^p)$, 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 $J^p$. Thus $\lambda \in \alpha(A)^p$ 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 $\sigma(A)^p \subseteq \sigma(A^p)$. This proves the lemma.

**Lemma C.9.8** Let $A$ be an $n \times n$ matrix and suppose $|\lambda| > \|A\|_2$. Then $(\lambda I - A)^{-1}$ exists.

**Proof:** Suppose $(\lambda I - A)x = 0$ where $x \neq 0$. Then

$$|\lambda| \|x\|_2 = \|Ax\|_2 \leq \|A\| \|x\|_2 < |\lambda| \|x\|_2,$$

a contradiction. Therefore, $(\lambda I - A)$ is one to one and this proves the lemma.

The main result is the following theorem due to Gelfand in 1941.

**Theorem C.9.9** Let $A$ be an $n \times n$ matrix. Then for any $\|\cdot\|$ defined on $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$

$$\rho(A) = \lim_{p \to \infty} \|A^p\|^{1/p}.$$

**Proof:** If $\lambda \in \sigma(A)$, then by Lemma C.9.7 $\lambda^p \in \sigma(A^p)$ and so by Lemma C.9.8, it follows that

$$|\lambda|^p \leq \|A^p\|$$

and so $|\lambda| \leq \|A^p\|^{1/p}$. Since this holds for every $\lambda \in \sigma(A)$, it follows that for each $p$,

$$\rho(A) \leq \|A^p\|^{1/p}.$$

Now using Lemma C.9.6,

$$\rho(A) \geq \limsup_{p \to \infty} \|A^p\|^{1/p} \geq \liminf_{p \to \infty} \|A^p\|^{1/p} \geq \rho(A)$$

which proves the theorem.

### C.10 Convergence For Iterative Methods

Here I will give a proof of Theorem 31.1.8 based on Gelfand’s theorem.

**Lemma C.10.1** Suppose $T : \mathbb{F}^n \to \mathbb{F}^n$. Here $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. Also suppose

$$|Tx - Ty| \leq r |x - y|$$

for some $r \in (0, 1)$. Then there exists a unique fixed point, $x \in \mathbb{F}^n$ such that

$$Tx = x.$$

Letting $x^1 \in \mathbb{F}^n$, this fixed point, $x$, is the limit of the sequence of iterates,

$$x^1, T x^1, T^2 x^1, \ldots$$

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.

$$|x^1 - x| \leq \frac{1}{1 - r} |x^1 - Tx^1|.$$
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
\[
|T^k x_1 - T^{k-1} x_1| \leq r^{k-1} |T x_1 - x_1|.
\] (3.32)
Then
\[
|T^{k+1} x_1 - T^k x_1| \leq r |T^k x_1 - T^{k-1} x_1|
\leq rr^{k-1} |T x_1 - x_1| = r^k |T x_1 - x_1|.
\]
By induction, this shows that for all \( k \geq 2 \), (3.32) is valid. Now let \( k > l \geq N \).
\[
|x_1 - T^k x_1| = \sum_{j=1}^{k-1} (T^{j+1} x_1 - T^j x_1)
\leq \sum_{j=l}^{k-1} |T^{j+1} x_1 - T^j x_1|
\leq \sum_{j=N}^{k-1} r^j |T x_1 - x_1| \leq |T x_1 - x_1| \frac{r^N}{1-r}
\]
which converges to 0 as \( N \to \infty \). Therefore, this is a Cauchy sequence so it must converge to \( x \in \mathbb{F}^n \). Then
\[
x = \lim_{k \to \infty} T^k x_1 = \lim_{k \to \infty} T^{k+1} x_1 = T \lim_{k \to \infty} T^k x_1 = T x.
\]
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. This proves the lemma.

The following corollary is what will be used to prove the convergence condition for the various iterative procedures.

**Corollary C.10.2** Suppose \( T : \mathbb{F}^n \to \mathbb{F}^n \), for some constant \( C \)
\[
|T x - T y| \leq C |x - y|,
\]
for all \( x, y \in \mathbb{F}^n \), and for some \( N \in \mathbb{N} \),
\[
|T^N x - T^N y| \leq r |x - y|,
\]
for all \( x, y \in \mathbb{F}^n \) 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 C.10.1 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.

It remains to consider the convergence of the sequence. Without loss of generality, it can be assumed \( C \geq 1 \). Then if \( r \leq N - 1 \),

\[
|T^r x - T^s x| \leq C^N |x - y|
\]

for all \( x, y \in \mathbb{F}^n \). By Lemma C.10.1 there exists \( K \) such that if \( k, l \geq K \), then

\[
|T^{kN} x^1 - T^{lN} x^1| < \eta = \frac{\varepsilon}{2C^N}
\]

and also \( K \) is large enough that

\[
2r^K C^N \frac{|T^N x^1 - x^1|}{1 - r} < \frac{\varepsilon}{2}
\]

(3.35)

Now let \( p, q > KN \) and define \( k_p, k_q, r_p, \) and \( r_q \) by

\[
p = k_p N + r_p, \quad q = k_q N + r_q, \quad 0 \leq r_p, r_q < N.
\]

Then both \( k_p \) and \( k_q \) are larger than \( K \). Therefore, from (3.33) and (3.35),

\[
|T^p x^1 - T^q x^1| = |T^{r_p} T^{k_p N} x^1 - T^{r_q} T^{k_q N} x^1|
\]

\[
\leq |T^{k_p N} T^{r_p} x^1 - T^{k_q N} T^{r_q} x^1| + |T^{r_p} T^{k_p N} x^1 - T^{r_q} T^{k_q N} x^1|
\]

\[
\leq r_{k_p} |T^{r_p} x^1 - T^{r_q} x^1| + C^N |T^{k_p N} x^1 - T^{k_q N} x^1|
\]

\[
\leq r^K \left( \frac{|T^{r_p} x^1 - T^{r_q} x^1|}{1 - r_p} + \frac{|T^{r_q} x^1 - T^{r_q} x^1|}{1 - r_q} \right) + C^N \eta
\]

\[
\leq r^K \left( C^N |x^1 - x| + C^N |x^1 - x| \right) + C^N \eta
\]

\[
\leq 2r^K C^N \frac{|T^{N} x^1 - x^1|}{1 - r} + C^N \eta < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\]

This shows \( \{ T^{k} x^1 \} \) is a Cauchy sequence and since a subsequence converges to \( x \), it follows this sequence also must converge to \( x \). Here is why. Let \( \varepsilon > 0 \) be given. There exists \( M \) such that if \( k, l > M \), then

\[
|T^k x^1 - T^l x^1| < \frac{\varepsilon}{2}.
\]

Now let \( k > M \). Then let \( l > M \) and also be large enough that

\[
|T^{lN} x^1 - x| < \frac{\varepsilon}{2}.
\]

Then

\[
|T^k x^1 - x| \leq |T^k x^1 - T^{lN} x^1| + |T^{lN} x^1 - x|
\]

\[
\leq |T^k x^1 - T^{lN} x^1| + \frac{\varepsilon}{2}
\]

\[
< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\]

(3.36)

This proves the corollary.
Theorem C.10.3 Suppose for any norm on $L(\mathbb{F}^n, \mathbb{F}^n)$, $\rho(B^{-1}C) < 1$. Then the iterates in (31.7) converge to the unique solution of (31.6).

Proof: Consider the iterates in (31.7). Let $T x = B^{-1}C x + b$. Then

$$|T^k x - T^k y| = \left| (B^{-1}C)^k x - (B^{-1}C)^k y \right| \leq \|(B^{-1}C)^k\| |x - y|.$$

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 $\mathbb{F}^n$. Since $\rho(B^{-1}C) = \lim_{p \to \infty} \|(B^{-1}C)^p\|^{1/p} < 1$, it follows there exists $N$ such that if $k \geq N$, then for some $r^{1/k} < 1$,

$$\|(B^{-1}C)^k\|^{1/k} < r^{1/k} < 1.$$

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 C.10.2 applies and gives the conclusion of this theorem.
Bibliography


Index

$C^1$, 305
$C^k$, 305
$\Delta$, 422
$\nabla^2$, 422
$\sigma(A)$, 611

Abel’s formula, 75, 89
adjoint, 160
adjugate, 65, 84
algebraic multiplicity, 519
angle between vectors, 132
angular velocity, 148
arc length, 258
area of a parallelogram, 143
arithmetic mean, 358
augmented matrix, 25
back substitution, 24
balance of momentum, 433
basic variables, 31
basis, 604
bezier curves, 242
bounded, 223
Cartesian coordinates, 14
Cauchy Schwartz inequality, 130
Cauchy Schwarz, 115
Cauchy Schwarz inequality, 139, 631
Cauchy sequence, 225, 326
Cauchy sequence, 225
Cauchy stress, 435
Cavendish, 282
Cayley Hamilton theorem, 87
center of mass, 397
central force, 269
central force field, 281
centrifugal acceleration, 250
centripetal acceleration, 250
centripetal force, 281
chain rule, 311
change of variables formula, 383
characteristic equation, 183
characteristic polynomial, 86
characteristic value, 182
Christoffel symbols, 469
circulation density, 459
classical adjoint, 65
closed set, 117
coefficient of thermal conductivity, 337
cofactor, 58, 60, 82
cofactor matrix, 60
column vector, 42
compact, 228
companion matrix, 573
complement, 117
component, 123, 141
components of a matrix, 40
conformable, 44
conservation of linear momentum, 248
conservation of mass, 433
conservative, 453
constitutive laws, 438
contented set, 485
continuous function, 212
converge, 225
Coordinates, 13
Coriolis acceleration, 250
Coriolis acceleration
earth, 252
Coriolis force, 250, 281
Cramer’s rule, 68, 84
cross product
  general curvilinear coordinates, 475
curl, 421
curl
  general curvilinear coordinates, 475
curvature, 266
curvilinear coordinates, 467
cycloid, 459
defective, 520
defective eigenvalue, 520
deformation gradient, 434
dependent, 98
derivative, 303
derivative of a function, 233

determinant, 57, 77
    Laplace expansion, 60
    product, 80
    product of matrices, 62
    transpose, 79
diagonalizable, 197, 529, 622
diameter, 223
difference quotient, 233
differentiable, 301
differentiable matrix, 238
differential equations, 466
differential equations, 466
differential equations, 466
dimension of vector space, 606
direct sum, 613
directed line segment, 17
direction vector, 17
directrix, 141
distance formula, 114
divergence, 421
divergence, 471
divergence theorem.
    general curvilinear coordinates, 472
divergence theorem, 426
    Dolittle's method, 513
dominant eigenvalue, 559
donut, 407
dot product, 129
dual basis, 162
dual basis, 462
echelon form, 26
eigenspace, 184, 611
eigenvalue, 182, 357, 611
eigenvector, 182
Einstein summation convention, 152
elementary operations, 23
entries of a matrix, 40
equality of mixed partial derivatives, 297
equivalence class, 620
equivalence of norms, 633
equivalence relation, 620
Euler angles, 592
Eulerian coordinates, 433
exchange theorem, 100

Fibonacci sequence, 225
Fick's law, 337, 443
focus, 120
force, 122
force field, 261, 281
Foucault pendulum, 252

Fourier law of heat conduction, 337
Fredholm alternative, 161
free variables, 31
fundamental theorem line integrals, 453
fundamental theorem of algebra, 601

Gauss Elimination, 33
Gauss elimination, 26
Gauss Jordan method for inverses, 51
Gauss Seidel method, 552
Gauss's theorem, 426
general solution, 177
genralized eigenspace, 611
gometric mean, 358
gometric multiplicity, 519
Gerschgorin's theorem, 527
gradient, 295
gradient
    contravariant components, 471
covariant components, 470
gradient vector, 336
Gram Schmidt process, 158
Grammian, 190
grids, 477

harmonic, 298
heat equation, 298
Heine Borel, 271
Heine Borel theorem, 228
Hermitian, 195
Hessian matrix, 345
Holder's inequality, 635
homotopy method, 324

inconsistent, 30
inner product, 129
intercepts, 204
interior point, 117
inverses and determinants, 67, 83
invertible, 49

Jacobi method, 550
Jacobian determinant, 383
Jordan block, 624
Jordan content, 485
Jordan set, 485
joule, 135

Kepler's first law, 282
Kepler's laws, 281
Kepler's third law, 285
kilogram, 147
INDEX

kinetic energy, 247
kinetic energy, 465
Kronecker delta, 152

Lagrange multipliers, 354
Lagrangian, 319
Lagrangian coordinates, 433
Lagrangian formalism, 466
Laplace expansion, 60, 82
Laplacian, 298
Laplaceian
  general curvilinear coordinates, 472
leading entry, 26
least squares regression, 299
Lebesgue number, 228
Lebesgue’s theorem, 490
length of smooth curve, 259
limit of a function, 213, 231
limit point, 121, 291
line integral, 262
linear combination, 80, 94
linear momentum, 247
linear transformation, 170, 303, 608
linearly independent, 98, 604
Lipschitz, 217, 218
lizards
  surface area, 404
local extremum, 341
local maximum, 341
local minimum, 341
lower sum, 370, 478

main diagonal, 61
mass ballance, 433
material coordinates, 433
matrix, 39
  inverse, 49
  left inverse, 84
  lower triangular, 61, 84
  non defective, 195
  normal, 195
  right inverse, 84
  self adjoint, 529, 548, 597
  symmetric, 529, 597
  upper triangular, 61, 84
matrix multiplication
  entries, 45
  properties, 47
matrix of linear transformation, 618
matrix transpose, 47
matrix transpose properties, 48

metric tensor, 163, 190
metric tensor, 470
migration matrix, 524
minimal polynomial, 610
minor, 58, 60, 82
moment of a force, 147
monic polynomial, 609
motion, 433
moving coordinate system, 239, 249
  acceleration, 250
multi-index, 212
Navier, 443
Newton, 124
  second law, 243
Newton Raphson method, 321
Newton’s method, 322
Newton’s second law, 465
nilpotent, 73, 617
nondefective eigenvalue, 520
one to one, 170
onto, 170
open cover, 228
open set, 117
operator norm, 326
orientable, 452
orientation, 260
oriented curve, 260
origin, 13
orthogonal matrix, 72, 533
orthonormal, 157, 534
osculating plane, 266

parameter, 16, 17
parametric equation, 16
parametrization, 258
partial derivative, 294
permutation symbol, 152
perpendicular, 133
Piola Kirchhoff stress, 437
pivot, 31
pivot column, 26, 501
pivot position, 26
precession of a top, 395
principal normal, 266
principle directions, 523
product of matrices, 44
product rule
  cross product, 236
dot product, 236
matrices, 238
radius of curvature, 266
rank, 503
rank of a matrix, 84, 503
real Schur form, 194
recurrence relation, 224
recursively defined sequence, 224
refinement of grids, 477
regression line, 160
resultant, 124
Riemann criterion, 479
Riemann integral, 479
rot, 421
row operations, 62, 499
row reduced echelon form, 500
row reduction algorithm, 27
row vector, 42
saddle point, 343
scalar field, 421
scalar multiplication, 15
scalar potential, 453
scalar product, 129
scalars, 15, 39, 92
second derivative test, 347
self adjoint, 195
sequences, 224
sequential compactness, 271
sequentially compact, 226
similar matrices, 620
similarity transformation, 620
simultaneous corrections, 550
skew symmetric, 48
smooth curve, 258
solution set, 23
spacial coordinates, 433
span, 80, 97
spanning set, 98
spectral radius, 637
spectrum, 182
speed, 125
spherical coordinates, 313
spherical coordinates, 463
standard matrix, 303
Stokes, 443
strictly upper triangular, 624
subspace, 97, 603
symmetric, 48
symmetric form of a line, 18
symmetric matrix, 535
torque vector, 147