# Contents

## I Introduction

### 11

## II Vectors, Vector Products, Lines

### 17

#### 1.1 \( \mathbb{R}^n \) Ordered \( n \) -tuples

#### 1.2 Vectors And Algebra In \( \mathbb{R}^n \)

#### 1.3 Geometric Meaning Of Vectors

#### 1.4 Geometric Meaning Of Vector Addition

#### 1.5 Distance Between Points In \( \mathbb{R}^n \) Length Of A Vector

#### 1.6 Geometric Meaning Of Scalar Multiplication

#### 1.7 Unit Vectors (Direction Vectors)

#### 1.8 Parametric Lines

#### 1.9 Vectors And Physics

#### 1.10 The Special Unit Vectors \( i, j, k \)

### 37

#### 2.1 The Dot Product

##### 2.1.1 Definition In terms Of Coordinates

##### 2.1.2 The Geometric Meaning Of The Dot Product, The Included Angle

##### 2.1.3 The Cauchy Schwarz Inequality

##### 2.1.4 The Triangle Inequality

##### 2.1.5 Direction Cosines

##### 2.1.6 Work

##### 2.1.7 Projections

#### 2.2 The Cross Product

##### 2.2.1 Cross Product, Geometric Description

##### 2.2.2 The Coordinate Description Of The Cross Product

##### 2.2.3 A Physical Application

##### 2.2.4 The Box Product, Triple Product

##### 2.2.5 A Proof Of The Distributive Law For The Cross Product

##### 2.2.6 Torque, Moment Of A Force

##### 2.2.7 Angular Velocity

##### 2.2.8 Center Of Mass

#### 2.3 Further Explanations

##### 2.3.1 The Distributive Law For The Cross Product

##### 2.3.2 Vector Identities And Notation

### 61

#### 2.4 Exercises With Answers
III Planes And Systems Of Equations

3 Planes

3.1 Finding Planes
3.1.1 The Cartesian Equation Of A Plane
3.1.2 The Plane Which Contains Three Points
3.1.3 The Cartesian Equation For Lines
3.1.4 Parametric Equation For A Plane
3.1.5 Triple Vector Product
3.1.6 The Angle Between Two Planes
3.1.7 Intercepts Of A Plane
3.1.8 Distance Between A Point And A Plane Or A Point And A Line

4 Systems Of Linear Equations

4.1 Systems Of Equations, Geometric Interpretations
4.2 Systems Of Equations, Algebraic Procedures
4.2.1 Elementary Operations
4.2.2 Gauss Elimination, Row Echelon Form

5 Matrices

5.1 Matrix Operations And Algebra
5.1.1 Addition And Scalar Multiplication Of Matrices
5.1.2 Multiplication Of Matrices
5.1.3 The $ij^{th}$ Entry Of A Product
5.1.4 Graphs And Digraphs
5.1.5 Properties Of Matrix Multiplication
5.1.6 The Transpose
5.1.7 The Identity And Inverses

5.2 Finding The Inverse Of A Matrix, Gauss Jordan Method
5.3 Systems Of Equations And Matrices
5.4 Elementary Matrices
5.5 Finding Linear Relationships Between Vectors
5.5.1 The Great And Glorious Lemma On Row Operations
5.5.2 Theory Of Row Reduced Echelon Form
5.6 Block Multiplication Of Matrices
5.7 Exercises With Answers
5.8 The Rank Of A Matrix
5.9 Exercises With Answers

6 The LU and QR Factorization

6.0.1 Definition Of An LU Factorization
6.0.2 Finding An LU Factorization By Inspection
6.0.3 Using Multipliers To Find An LU Factorization
6.0.4 Solving Systems Using A LU Factorization
6.1 The QR Factorization

7 Determinants

7.1 Basic Techniques And Properties
7.1.1 Cofactors And $2 \times 2$ Determinants
7.1.2 The Determinant Of A Triangular Matrix
7.1.3 Properties Of Determinants
7.1.4 Finding Determinants Using Row Operations
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CONTENTS</strong></td>
<td>5</td>
</tr>
<tr>
<td>7.1.5 A Formula For The Inverse</td>
<td>167</td>
</tr>
<tr>
<td>7.1.6 Cramer's Rule</td>
<td>171</td>
</tr>
<tr>
<td><strong>8</strong> The Mathematical Theory Of Determinants</td>
<td>173</td>
</tr>
<tr>
<td>8.0.7 The Function sgn</td>
<td>173</td>
</tr>
<tr>
<td>8.1 The Determinant</td>
<td>175</td>
</tr>
<tr>
<td>8.1.1 The Definition</td>
<td>175</td>
</tr>
<tr>
<td>8.1.2 Permuting Rows Or Columns</td>
<td>175</td>
</tr>
<tr>
<td>8.1.3 A Symmetric Definition</td>
<td>177</td>
</tr>
<tr>
<td>8.1.4 The Alternating Property Of The Determinant</td>
<td>177</td>
</tr>
<tr>
<td>8.1.5 Linear Combinations And Determinants</td>
<td>178</td>
</tr>
<tr>
<td>8.1.6 The Determinant Of A Product</td>
<td>179</td>
</tr>
<tr>
<td>8.1.7 Cofactor Expansions</td>
<td>181</td>
</tr>
<tr>
<td>8.1.8 Formula For The Inverse</td>
<td>181</td>
</tr>
<tr>
<td>8.1.9 Cramer's Rule</td>
<td>182</td>
</tr>
<tr>
<td>8.1.10 Upper Triangular Matrices</td>
<td>183</td>
</tr>
<tr>
<td><strong>S.2</strong> The Cayley Hamilton Theorem</td>
<td>183</td>
</tr>
<tr>
<td><strong>S.3</strong> Exercises With Answers</td>
<td>185</td>
</tr>
<tr>
<td><strong>9</strong> Vector Spaces Subspaces Bases</td>
<td>189</td>
</tr>
<tr>
<td>9.0.1 What Is A Vector Space?</td>
<td>189</td>
</tr>
<tr>
<td>9.0.2 What Is A Subspace?</td>
<td>191</td>
</tr>
<tr>
<td>9.0.3 What Is A Span?</td>
<td>194</td>
</tr>
<tr>
<td>9.1 Linear Independence, Bases, Dimension</td>
<td>195</td>
</tr>
<tr>
<td>9.1.1 Linear Independence For Spaces Of Functions</td>
<td>202</td>
</tr>
<tr>
<td>9.1.2 Row Space, Column Space, And Null Space</td>
<td>203</td>
</tr>
<tr>
<td><strong>9.2</strong> Row Column And Determinant Rank</td>
<td>208</td>
</tr>
<tr>
<td><strong>10</strong> Linear Transformations</td>
<td>211</td>
</tr>
<tr>
<td>10.0.1 Outcomes</td>
<td>211</td>
</tr>
<tr>
<td>10.1 What Is A Linear Transformation?</td>
<td>211</td>
</tr>
<tr>
<td>10.2 Linear Transformations</td>
<td>212</td>
</tr>
<tr>
<td>10.3 Constructing The Matrix Of A Linear Transformation</td>
<td>214</td>
</tr>
<tr>
<td>10.3.1 Rotations in ( \mathbb{R}^2 )</td>
<td>215</td>
</tr>
<tr>
<td>10.3.2 Projections</td>
<td>217</td>
</tr>
<tr>
<td>10.3.3 Algebraically Defined Linear Transformations</td>
<td>218</td>
</tr>
<tr>
<td>10.3.4 Matrices Which Are One To One Or Onto</td>
<td>219</td>
</tr>
<tr>
<td>10.3.5 Similarity</td>
<td>221</td>
</tr>
<tr>
<td>10.3.6 The General Solution Of A Linear System</td>
<td>221</td>
</tr>
<tr>
<td><strong>10.4</strong> Exercises With Answers</td>
<td>224</td>
</tr>
<tr>
<td><strong>11</strong> Eigenvalues And Eigenvectors Of A Matrix</td>
<td>227</td>
</tr>
<tr>
<td>11.1 Definition Of Eigenvectors And Eigenvalues</td>
<td>227</td>
</tr>
<tr>
<td>11.2 Finding Eigenvectors And Eigenvalues</td>
<td>229</td>
</tr>
<tr>
<td>11.3 A Warning</td>
<td>232</td>
</tr>
<tr>
<td>11.4 Some Useful Facts</td>
<td>234</td>
</tr>
<tr>
<td>11.5 Defective And Nondefective Matrices</td>
<td>234</td>
</tr>
<tr>
<td>11.6 Diagonalization</td>
<td>238</td>
</tr>
<tr>
<td>11.7 The Matrix Exponential</td>
<td>241</td>
</tr>
<tr>
<td>11.8 Migration Matrices</td>
<td>243</td>
</tr>
<tr>
<td>11.9 Complex Eigenvalues</td>
<td>248</td>
</tr>
<tr>
<td>11.10 The Estimation Of Eigenvalues</td>
<td>249</td>
</tr>
</tbody>
</table>
12 Curves In Space
12.1 Limits Of A Vector Valued Function Of One Variable 259
12.2 The Derivative And Integra 261
12.2.1 Arc Length 263
12.2.2 Geometric And Physical Significance Of The Derivative 265
12.2.3 Differentiation Rules 266
12.2.4 Leibniz's Notation 269
12.2.5 Motion Of A Projectile (Idealized) 269
12.3 Exercises With Answers 271

13 Newton's Laws Of Motion 273
13.1 Kinetic Energy 277
13.2 Impulse And Momentum 278
13.3 Conservation Of Momentum 278
13.4 Exercises With Answers 279

14 Physics Of Curvilinear Motion 281
14.1 The Acceleration In Terms Of The Unit Tangent And Normal 281
14.2 The Circle Of Curvature 287
14.3 Independence Of Parametrization 292
14.3.1 Hard Calculus 293
14.3.2 Independence Of Parametrization 297
14.3.3 Product Rule For Matrices 299
14.4 Moving Coordinate Systems 299

IV Functions Of Many Variables 303

15 Functions Of Many Variables 305
15.1 The Graph Of A Function Of Two Variables 305
15.2 The Domain Of A Function 307
15.3 Quadric Surfaces 307
15.4 Open And Closed Sets 311
15.5 Continuous Functions 315
15.6 Sufficient Conditions For Continuity 316
15.7 Properties Of Continuous Functions 317

16 Limits Of A Function 319
16.1 The Limit Of A Function Of Many Variables 319
16.2 The Directional Derivative And Partial Derivatives 323
16.2.1 The Directional Derivative 323
16.2.2 Partial Derivatives 324
16.2.3 Mixed Partial Derivatives 328
16.3 Some Fundamentals 330
16.3.1 The Nested Interval Lemma 333
16.3.2 The Extreme Value Theorem 333
16.3.3 Sequences And Completeness 335
16.3.4 Continuity And The Limit Of A Sequence 338
CONTENTS

V Differentiability

17 Differentiability

17.1 The Definition of Differentiability 341
17.2 $C^1$ Functions and Differentiability 343
17.3 The Directional Derivative
  17.3.1 Separable Differential Equations 346
17.4 Exercises With Answers 349
17.4.1 A Heat Seaking Particle 350
17.5 The Chain Rule 350
17.5.1 Related Rates Problems 353
17.6 Taylor's Formula An Application Of Chain Rule 355
17.7 Normal Vectors And Tangent Planes 358

18 Extrema Of Functions Of Several Variables

18.1 Local Extrema 362
18.2 The Second Derivative Test
  18.2.1 Functions Of Two Variables 364
  18.2.2 Functions Of Many Variables 365
18.3 Lagrange Multipliers 368
18.4 Finding Function From Its Gradient 374
18.5 Exercises With Answers 376

19 The Derivative Of Vector Valued Functions, What Is The Derivative?

19.1 $C^1$ Functions 381
19.2 The Chain Rule
  19.2.1 The Chain Rule For Functions Of One Variables 385
  19.2.2 The Chain Rule For Functions Of Many Variables 385
  19.2.3 The Derivative Of The Inverse Function 389
19.3 Proof Of The Chain Rule 393
19.4 Proof Of The Second Derivative Test 394

20 Implicit Function Theorem

20.1 The Method Of Lagrange Multipliers 403
20.2 The Local Structure Of $C^1$ Mappings 404

VI Multiple Integrals

21 The Riemann Integral On $\mathbb{R}^n$ 409

21.1 Methods For Double Integrals
  21.1.1 Density Mass And Center Of Mass 416
21.2 Double Integrals In Polar Coordinates 417
21.3 A Few Applications 419
21.4 Surface Area And Integrals 422
21.5 Methods For Triple Integrals
  21.5.1 Definition Of The Integral 426
  21.5.2 Iterated Integrals 428
  21.5.3 Mass And Density 431
21.6 Exercises With Answers 433
## CONTENTS

### 22 The Integral In Other Coordinates

- **22.1 Cylindrical And Spherical Coordinates**
  - 22.1.1 Geometric Description
  - 22.1.2 Volume And Integrals In Cylindrical Coordinates
  - 22.1.3 Volume And Integrals In Spherical Coordinates

- **22.2 Exercises**
  - 22.3 The General Procedure

### 22.4 Exercises With Answers

- **22.5 The Moment Of Inertia**
  - 22.5.1 The Spinning Top
  - 22.5.2 Kinetic Energy

- **22.6 Finding The Moment Of Inertia And Center Of Mass**

### 22.7 Exercises With Answers

### VII Line Integrals

#### 23 Line Integrals

- **23.0.1 Orientations And Smooth Curves**
- **23.0.2 The Integral Of A Function Defined On A Smooth Curve**
- **23.0.3 Vector Fields**
- **23.0.4 Line Integrals And Work**
- **23.0.5 Another Notation For Line Integrals**

- **23.1 Exercises With Answers**

- **23.2 Conservative Vector Fields, Path Independence**
  - 23.2.1 Finding The Scalar Potential, (Recover The Function From Its Gradient)
  - 23.2.2 Terminology

- **23.3 Divergence, Gradient, Curl**
  - 23.3.1 Divergence
  - 23.3.2 Gradient
  - 23.3.3 Curl
  - 23.3.4 Identities Involving Curl Divergence And Gradient
  - 23.3.5 Vector Potentials

### VIII Green’s Theorem, Integrals On Surfaces

#### 24 Green’s Theorem

- **24.1 An Alternative Explanation Of Green’s Theorem**
- **24.2 Area And Green’s Theorem**

#### 25 The Integral On Two Dimensional Surfaces In $\mathbb{R}^3$

- **25.1 Parametrically Defined Surfaces**
- **25.2 The Two Dimensional Area In $\mathbb{R}^3$**
  - 25.2.1 Surfaces Of The Form $z = f(x, y)$

- **26.4 Exercises With Answers**
## The Divergence Theorem And Stoke's Theorem

### 26 Divergence And Curl

- **26.1 Divergence Of A Vector Field**
- **26.2 Curl Of A Vector Field**

### 27 The Divergence Theorem

- **27.0.1 Green's Theorem, A Review**
- **27.0.2 Coordinate Free Concept Of Divergence, Flux Density**
- **27.1 The Weak Maximum Principle**
- **27.2 Some Applications Of The Divergence Theorem**
  - **27.2.1 Hydrostatic Pressure**
  - **27.2.2 Archimedes Law Of Buoyancy**
  - **27.2.3 Equations Of Heat And Diffusion**
  - **27.2.4 Balance Of Mass**
  - **27.2.5 Balance Of Momentum**
  - **27.2.6 Frame Indifference**
  - **27.2.7 Bernoulli's Principle**
  - **27.2.8 The Wave Equation**
  - **27.2.9 A Negative Observation**
  - **27.2.10 Electrostatics**

### 28 Stokes And Green's Theorems

- **28.1 Green's Theorem**
- **28.2 Exercises**
- **28.3 Stoke's Theorem From Green's Theorem**
  - **28.3.1 The Normal And The Orientation**
  - **28.3.2 The Mobeus Band**
  - **28.3.3 The Meaning Of The Curl**
  - **28.3.4 Conservative Vector Fields**
  - **28.3.5 Some Terminology**
- **28.4 When Do You Use What?**
- **28.5 Maxwell's Equations And The Wave Equation**

## Some Iterative Techniques For Linear Algebra

### 29 Iterative Methods For Linear Systems

- **29.1 Jacobi Method**
- **29.2 Gauss Seidel Method**

### 30 Iterative Methods For Finding Eigenvalues

- **30.1 The Power Method For Eigenvalues**
  - **30.1.1 Rayleigh Quotient**
- **30.2 The Shifted Inverse Power Method**

## The Correct Version Of The Riemann Integral

### A The Theory Of The Riemann Integral

- **A.1 An Important Warning**
- **A.2 The Definition Of The Riemann Integral**
A.3 Basic Properties ......................................................... 596
A.4 Which Functions Are Integrable? ................................. 598
A.5 Iterated Integrals ......................................................... 607
A.6 The Change Of Variables formula ................................. 611
A.7 Some Observations ..................................................... 617

Copyright © 2005,
Part I

Introduction
These are the lecture notes for my section of Math 302. They are pretty much in the order of the syllabus for the course. You don’t need to read the starred sections and chapters and subsections. These are there to provide depth in the subject. To quote from the mission statement of BYU, “Depth comes when students realize the effect of rigorous, coherent, and progressively more sophisticated study. Depth helps students distinguish between what is fundamental and what is only peripheral; it requires focus, provides intense concentration. ...” To see clearly what is peripheral you need to read the fundamental and difficult concepts, most of which are presented in the starred sections. These are not always easy to read and I have indicated the most difficult with a picture of a dragon. Some are not much harder than what is presented in the course. A good example is the one which defines the derivative. If you don’t learn this material, you will have trouble understanding many fundamental topics. Some which come to mind are basic continuum mechanics (The deformation gradient is a derivative.) and Newton’s method for solving nonlinear systems of equations.(The entire method involves looking at the derivative and its inverse.) If you don’t want to learn anything more than what you will be tested on, then you can omit these sections. This is up to you. It is your choice.

A word about notation might help. Most of the linear algebra works in any field. Examples are the rational numbers, the integers modulo a prime number, the complex numbers, or the real numbers. Therefore, I will often write \( F \) to denote this field. If you don’t like this, just put in \( \mathbb{R} \) and you will be fine. This is the main one of interest. However, I at least want you to realize that everything holds for the complex numbers in addition to the reals. In many applications this is essential so it does not hurt to begin to realize this. Also, I will write vectors in terms of bold letters. Thus \( \mathbf{u} \) will denote a vector. Sometimes people write something like \( \vec{u} \) to indicate a vector. However, the bold face is the usual notation so I am using this in these notes. On the board, I will likely write the other notation. The norm or length of a vector is often written as \( ||\mathbf{u}|| \). I will usually write it as \( |\mathbf{u}| \). This is standard notation also although most books use the double bar notation. The notation I am using emphasizes that the norm is just like the absolute value which is an important connection to make. It also seems less cluttered. You need to understand that either notation means the same thing.

For a more substantial treatment of certain topics, there is a complete calculus book on my web page. There are significant generalizations which unify all the notions of volume into one beautiful theory. I have not pursued this topic in these notes but it is in the calculus book. There are other things also, especially all the one variable theory if you need a review.
Part II

Vectors, Vector Products, Lines
Chapter 1

Vectors And Points In $\mathbb{R}^n$

1.1 $\mathbb{R}^n$ Ordered $n-$ tuples

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, $\mathbf{x}$. 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 coordinate axis, the $x_i$ axis for short. The point $\mathbf{0} \equiv (0, \cdots, 0)$ 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 René 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.

### 1.2 Vectors And Algebra In \(\mathbb{R}^n\)

There are two algebraic operations done with points 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

\[
ax = a (x_1, \cdots, x_n) \equiv (ax_1, \cdots, ax_n).
\]

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

\[
x + y = (x_1, \cdots, x_n) + (y_1, \cdots, y_n)
\equiv (x_1 + y_1, \cdots, x_n + y_n)
\]

\(^1\)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.
1.3. GEOMETRIC MEANING OF VECTORS

An element of \( \mathbb{R}^n \), \( \mathbf{x} \equiv (x_1, \cdots, x_n) \) is often called a vector. The above definition is known as vector addition.

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

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

\[
\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}, \tag{1.3}
\]
the commutative law of addition,

\[
(\mathbf{v} + \mathbf{w}) + \mathbf{z} = \mathbf{v} + (\mathbf{w} + \mathbf{z}), \tag{1.4}
\]
the associative law for addition,

\[\mathbf{v} + \mathbf{0} = \mathbf{v}, \tag{1.5}\]
the existence of an additive identity,

\[\mathbf{v} + (-\mathbf{v}) = \mathbf{0}, \tag{1.6}\]
the existence of an additive inverse, Also

\[
\alpha (\mathbf{v} + \mathbf{w}) = \alpha \mathbf{v} + \alpha \mathbf{w}, \tag{1.7}
\]
\[
(\alpha + \beta) \mathbf{v} = \alpha \mathbf{v} + \beta \mathbf{v}, \tag{1.8}
\]
\[
\alpha (\beta \mathbf{v}) = \alpha \beta (\mathbf{v}), \tag{1.9}
\]
\[1 \mathbf{v} = \mathbf{v}. \tag{1.10}\]

In the above \( \mathbf{0} = (0, \cdots, 0) \).

You should verify these properties all hold. For example, consider \ref{1.7}

\[
\alpha (\mathbf{v} + \mathbf{w}) = \alpha (v_1 + w_1, \cdots, v_n + w_n)
\]
\[
= (\alpha (v_1 + w_1), \cdots, \alpha (v_n + w_n))
\]
\[
= (\alpha v_1 + \alpha w_1, \cdots, \alpha v_n + \alpha w_n)
\]
\[
= (\alpha v_1, \cdots, \alpha v_n) + (\alpha w_1, \cdots, \alpha w_n)
\]
\[
= \alpha \mathbf{v} + \alpha \mathbf{w}.
\]

As usual subtraction is defined as \( \mathbf{x} - \mathbf{y} \equiv \mathbf{x} + (-\mathbf{y}). \)

1.3 Geometric Meaning Of Vectors

**Definition 1.3.1** Let \( \mathbf{x} = (x_1, \cdots, x_n) \) be the coordinates of a point in \( \mathbb{R}^n \). Imagine an arrow with its tail at \( \mathbf{0} = (0, \cdots, 0) \) and its point at \( \mathbf{x} \) as shown in the following picture in the case of \( \mathbb{R}^3 \).

\[
(x_1, x_2, x_3) = \mathbf{x}
\]
Then this arrow is called the position vector of the point, \( \mathbf{x} \). Given two points, \( P, Q \) whose coordinates are \((p_1, \cdots, p_n)\) and \((q_1, \cdots, q_n)\) respectively, one can also determine the vector defined as follows. Also one can obtain a vector from a given two points.

\[
\overrightarrow{PQ} \equiv (q_1 - p_1, \cdots, q_n - p_n)
\]

Thus every point determines such a vector and conversely, every such vector (arrow) which has its tail at \( \mathbf{0} \) determines a point of \( \mathbb{R}^n \), namely the point of \( \mathbb{R}^n \) which coincides with the point of the vector.

Imagine taking the above position vector and moving it around, always keeping it pointing in the same direction as shown in the following picture.

After moving it around, it is regarded as the same vector because it points in the same direction and has the same length. Thus each of the arrows in the above picture is regarded as the same vector. The components of this vector are the numbers, \( x_1, \cdots, x_n \). You should think of these numbers as directions for obtaining an arrow. Starting at some point, \((a_1, a_2, \cdots, a_n)\) in \( \mathbb{R}^n \), you move to the point \((a_1 + x_1, \cdots, a_n)\) and from there to the point \((a_1 + x_1, a_2 + x_2, a_3 + x_3, \cdots, a_n)\) and then to \((a_1 + x_1, a_2 + x_2, a_3 + x_3, \cdots, a_n + x_n)\). The arrow having its tail at \((a_1, a_2, \cdots, a_n)\) and its point at \((a_1 + x_1, a_2 + x_2, \cdots, a_n + x_n)\) looks just like the arrow which has its tail at \( \mathbf{0} \) and its point at \((x_1, \cdots, x_n)\) so it is regarded as the same vector.

### 1.4 Geometric Meaning Of Vector Addition

It was explained earlier that an element of \( \mathbb{R}^n \) is an \( n \) tuple of numbers and it was also shown that this can be used to determine a point in three dimensional space in the case where \( n = 3 \) and in two dimensional space, in the case where \( n = 2 \). This point was specified relative to some coordinate axes.

Consider the case where \( n = 3 \) for now. If you draw an arrow from the point in three dimensional space determined by \((0,0,0)\) to the point \((a,b,c)\) with its tail sitting at the point \((0,0,0)\) and its point at the point \((a,b,c)\), this arrow is called the position vector of the point determined by \( \mathbf{u} \equiv (a,b,c) \). One way to get to this point is to start at \((0,0,0)\) and move in the direction of the \( x_1 \) axis to \((a,0,0)\) and then in the direction of the \( x_2 \) axis to \((a,b,0)\) and finally in the direction of the \( x_3 \) axis to \((a,b,c)\). It is evident that the same arrow (vector) would result if you began at the point, \( \mathbf{v} \equiv (d,e,f) \), moved in the direction of the \( x_1 \) axis to \((d+a,e,f)\), then in the direction of the \( x_2 \) axis to \((d+a,e+b,f)\), and finally in the \( x_3 \) direction to \((d+a,e+b,f+c)\) only this time, the arrow would have its tail sitting at the point determined by \( \mathbf{v} \equiv (d,e,f) \) and its point at \((d+a,e+b,f+c)\). It is said to be the same arrow (vector) because it will point in the same direction and have the same length. It is like you took an actual arrow, the sort of thing you shoot with a bow, and moved it from one location to another keeping it pointing the same direction. This is illustrated in the following picture in which \( \mathbf{v} + \mathbf{u} \) is illustrated. Note the parallelogram determined in the picture by the vectors \( \mathbf{u} \) and \( \mathbf{v} \).

\(^2\text{I will discuss how to define length later. For now, it is only necessary to observe that the length should be defined in such a way that it does not change when such motion takes place.}\)
1.5. DISTANCE BETWEEN POINTS IN $\mathbb{R}^N$ LENGTH OF A VECTOR

Thus the geometric significance of $(d, e, f) + (a, b, c) = (d + a, e + b, f + c)$ is this. You start with the position vector of the point $(d, e, f)$ and at its point, you place the vector determined by $(a, b, c)$ with its tail at $(d, e, f)$. Then the point of this last vector will be $(d + a, e + b, f + c)$. This is the geometric significance of vector addition. Also, as shown in the picture, $u + v$ is the directed diagonal of the parallelogram determined by the two vectors $u$ and $v$. A similar interpretation holds in $\mathbb{R}^n, n > 3$ but I can’t draw a picture in this case.

Since the convention is that identical arrows pointing in the same direction represent the same vector, the geometric significance of vector addition is as follows in any number of dimensions.

**Procedure 1.4.1** Let $u$ and $v$ be two vectors. Slide $v$ so that the tail of $v$ is on the point of $u$. Then draw the arrow which goes from the tail of $u$ to the point of the slid vector, $v$. This arrow represents the vector $u + v$.

Note that $P + P\overrightarrow{Q} = Q$.

1.5 Distance Between Points In $\mathbb{R}^n$ Length Of A Vector

How is distance between two points in $\mathbb{R}^n$ defined?

**Definition 1.5.1** Let $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$ be two points in $\mathbb{R}^n$. Then $|x - y|$ to indicates the distance between these points and is defined as

$$distance \text{ between } x \text{ and } y \equiv |x - y| \equiv \left( \sum_{k=1}^{n} |x_k - y_k|^2 \right)^{1/2}.$$
This is called the distance formula. Thus $|x| \equiv |x - 0|$. 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 means all points in $\mathbb{R}^n$ which are closer to $a$ than $r$. The length of a vector $x$ is the distance between $x$ and $0$.

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. Often people use two lines to denote this distance, $||x - y||$. However, I want to emphasize this is really just like the absolute value. Also, the notation I am using is fairly standard.

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. In other words, this distance defined above is the same as the distance of plane geometry in which the Pythagorean theorem holds.

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
1.5. DISTANCE BETWEEN POINTS IN $\mathbb{R}^N$ LENGTH OF A VECTOR

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 no problem with the formula for distance in any number of dimensions. Here is an example.

**Example 1.5.2** 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 1.5.3** 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
\[
x^2 - 2x + 14 + y^2 - 4y - 6z = x^2 + y^2 - 2y + 5 + z^2 - 4z
\]
which implies
\[
-2x + 14 - 4y - 6z = -2y + 5 - 4z
\]
and so
\[
2x + 2y + 2z = -9. \tag{1.11}
\]
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 \((1.11)\) holds.

There are certain properties of the distance which are obvious. Two of them which follow directly from the definition are
\[
|x - y| = |y - x|,
\]
\[
|x - y| \geq 0 \text{ and equals 0 only if } y = x.
\]
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. I will show you a proof of this pretty soon. This is usually stated as
\[
|x + y| \leq |x| + |y|.
\]
Here is a picture which illustrates the statement of this inequality in terms of geometry.

\[\text{1.6 Geometric Meaning Of Scalar Multiplication}\]

As discussed earlier, \(x = (x_1, x_2, x_3)\) determines a vector. You draw the line from \(0\) to \(x\) placing the point of the vector on \(x\). What is the length of this vector? The length of this vector is defined to equal \(|x|\) as in Definition \[\|\|\]. Thus the length of \(x\) equals \(\sqrt{x_1^2 + x_2^2 + x_3^2}\). When you multiply \(x\) by a scalar, \(\alpha\), you get \((\alpha x_1, \alpha x_2, \alpha x_3)\) and the length of this vector is defined as \(\sqrt{((\alpha x_1)^2 + (\alpha x_2)^2 + (\alpha x_3)^2)} = |\alpha| \sqrt{x_1^2 + x_2^2 + x_3^2}\). Thus the following holds.
\[
|\alpha x| = |\alpha| |x|.
\]
In other words, multiplication by a scalar magnifies the length of the vector. What about the direction? You should convince yourself by drawing a picture that if \(\alpha\) is negative, it causes the resulting vector to point in the opposite direction while if \(\alpha > 0\) it preserves the direction the vector points.
1.6. GEOMETRIC MEANING OF SCALAR MULTIPLICATION

You can think of vectors as quantities which have direction and magnitude, little arrows. Thus any two little arrows which have the same length and point in the same direction are considered to be the same vector even if their tails are at different points.

You can always slide such an arrow and place its tail at the origin. If the resulting point of the vector is \((a, b, c)\), it is clear the length of the little arrow is \(\sqrt{a^2 + b^2 + c^2}\). Geometrically, the way you add two geometric vectors is to place the tail of one on the point of the other and then to form the vector which results by starting with the tail of the first and ending with this point as illustrated in the following picture. Also when \((a, b, c)\) is referred to as a vector, you mean any of the arrows which have the same direction and magnitude as the position vector of this point. Geometrically, for \(u = (u_1, u_2, u_3)\), \(\alpha u\) is any of the little arrows which have the same direction and magnitude as \((\alpha u_1, \alpha u_2, \alpha u_3)\).

The following example is art which illustrates these definitions and conventions.

Exercise 1.6.1 Here is a picture of two vectors, \(u\) and \(v\).

Sketch a picture of \(u + v\), \(u - v\), and \(u + 2v\).

First here is a picture of \(u + v\). You first draw \(u\) and then at the point of \(u\) you place the tail of \(v\) as shown. Then \(u + v\) is the vector which results which is drawn in the following pretty picture.
Next consider $u - v$. This means $u + (-v)$. From the above geometric description of vector addition, $-v$ is the vector which has the same length but which points in the opposite direction to $v$. Here is a picture.

Finally consider the vector $u + 2v$. Here is a picture of this one also.

### 1.7 Unit Vectors (Direction Vectors)

Let $v$ be a vector,

$$v = (v_1, \ldots, v_n).$$

The **direction vector** for $v$ is defined as $v/|v|$. This vector points in the same direction as $v$ because it consists of the scalar, $1/|v|$ times $v$. This vector is called a **unit vector** because $|v/|v|| = |v|/|v| = 1$. That is, it has length equal to 1. The process of dividing a vector by its length is called **normalizing**. It provides you with a vector which has unit length and the same direction as the original vector. This is called the direction vector of the vector $v$.

**Example 1.7.1** Let $P = (1, 2, 1), Q = (2, -1, 3)$. Find the direction vector of $\overline{PQ}$.

The vector determined by the two points has coordinates $(1, -3, 2)$ and so the direction vector is the unit vector which has the same direction, namely

$$\frac{(1, -3, 2)}{\sqrt{1 + 9 + 4}} = \left(\frac{1}{14}\sqrt{14}, -\frac{3}{14}\sqrt{14}, \frac{1}{7}\sqrt{14}\right).$$
1.8. PARAMETRIC LINES

**Definition 1.7.2** Two vectors are parallel if they have either the same direction vector or one direction vector is \(-1\) times the other. They are opposite in direction if their direction vectors add to \(0\).

1.8 Parametric 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 \(l\),

\[
\begin{align*}
(m & = \frac{y_2 - y_1}{x_2 - x_1} = \frac{y - y_1}{x - x_1} \\
& \Rightarrow y - y_1 = m (x - x_1) = \frac{y_2 - y_1}{x_2 - x_1} (x - x_1). \quad (1.12)
\end{align*}
\]

Now consider points of the form

\[
(x, y) = (x_1, y_1) + t (x_2 - x_1, y_2 - y_1). \quad (1.13)
\]

Do these points satisfy the above equation of the line? Is

\[
y_1 + t (y_2 - y_1) - y_1 = \left( \frac{y_2 - y_1}{x_2 - x_1} \right) (x_1 + t (x_2 - x_1) - x_1)?
\]

Yes, this is so. Both sides equal \(t (y_2 - y_1)\). Conversely, if \((x, y)\) is a point which satisfies the equation, \(1.12\) does there exist a value of \(t\) such that this point is of the form \((x_1, y_1) + t (x_2 - x_1, y_2 - y_1)\)? If the point satisfies \(1.12\), it is of the form

\[
\begin{align*}
(x, y_1 + \frac{y_2 - y_1}{x_2 - x_1} (x_1 + t (x_2 - x_1) - x_1)).
\end{align*}
\]
Now let \( t = \frac{x - x_1}{x_2 - x_1} \) so 
\[
x = x_1 + t \left( x_2 - x_1 \right).
\]
Then in terms of \( t \), the above reduces to
\[
\left( x_1 + t \left( x_2 - x_1 \right), y_1 + \left( \frac{y_2 - y_1}{x_2 - x_1} \right) t \left( x_2 - x_1 \right) \right) = (x_1, y_1) + t \left( x_2 - x_1, y_2 - y_1 \right).
\]
It follows the set of points in \( \mathbb{R}^2 \) obtained from \( \ref{eq:1.12} \) and \( \ref{eq:1.13} \) are the same. The following is the definition of a line in \( \mathbb{R}^n \).

**Definition 1.8.1** A line in \( \mathbb{R}^n \) containing the two different points, \( \mathbf{x}^1 \) and \( \mathbf{x}^2 \) is the collection of points of the form
\[
\mathbf{x} = \mathbf{x}^1 + t \left( \mathbf{x}^2 - \mathbf{x}^1 \right)
\]
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. From now on, you should think of lines in this way. Forget about the stupid special case in \( \mathbb{R}^2 \) which you had drilled in to your head in high school. The concept of a line is really very simple and it holds in any number of dimensions, not just in two dimensions. It is given in the above definition.

**Lemma 1.8.2** Let \( \mathbf{a}, \mathbf{b} \in \mathbb{R}^n \) with \( \mathbf{a} \neq 0 \). 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 \( t \mathbf{a} + \mathbf{b} = \mathbf{x}^1 + t \left( \mathbf{x}^2 - \mathbf{x}^1 \right) \) 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 1.8.3** The vector \( \mathbf{a} \) in the above lemma is called a **direction vector** for the line.

Direction vectors are what it is all about, not slope. Slope is fine in two dimensions but we live in three dimensions. Slope is a trivial and stupid concept designed mainly to give children something to do in high school. The correct and worthwhile notion is that of direction vector. This is a new concept. Do not try to fit it in to the stuff you saw earlier. Do not try to put the new wine in the old bottles, to quote the scripture. It only creates confusion and you do not need that.

**Example 1.8.4** Find the line through \( (1, 2) \) and \( (4, 7) \).

A vector equation of this line is \( (x, y) = (1, 2) + t (3, 5) \). Now if you want to get the equation in the form you are used to seeing in high school,
\[
x = 1 + 3t, \ y = 2 + 5t
\]
Solving the first one for \( t \), you get \( t = (x - 1) / 3 \) and now plugging this in to the second yields,
\[
y = 2 + 5 \left( \frac{x - 1}{3} \right)
\]
so \( y - 2 = \frac{5}{3} (x - 1) \) which is the usual point slope form for this line.

Now that you know about lines, it is possible to give a more analytical description of a vector as a directed line segment.
Definition 1.8.5 Let \( p \) and \( q \) be two points in \( \mathbb{R}^n \), \( p \neq q \). The directed line segment from \( p \) to \( q \), denoted by \( \overrightarrow{pq} \), is defined to be the collection of points,

\[
x = p + t(q - p), \quad t \in [0, 1]
\]

with the direction corresponding to increasing \( t \). In the definition, when \( t = 0 \), the point \( p \) is obtained and as \( t \) increases other points on this line segment are obtained until when \( t = 1 \), you get the point, \( q \). This is what is meant by saying the direction corresponds to increasing \( t \).

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

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

Example 1.8.6 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

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

The vector \((1, -6, 6)\) is obtained by \((2, -4, 6) - (1, 2, 0)\) as indicated above.

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 is 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 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 1.8.7 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

\[(x, y, z) = (1, 2, 0) + t(1, 2, 1), \quad t \in \mathbb{R}. \quad (1.14)
\]

Sometimes people elect to write a line like the above in the form

\[x = 1 + t, \quad y = 2 + 2t, \quad z = t, \quad t \in \mathbb{R}. \quad (1.15)
\]

This is a set of scalar parametric equations which amounts to the same thing as \((1.14)\).

There is one other form for a line which is sometimes considered useful. It is the so called symmetric form. Consider the line of \((1.15)\). You can solve for the parameter, \( t \) to write

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

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

This is the symmetric form of the line. Later, it will become clear that this expresses the line as the intersection of two planes but this is not important at this time.

**Example 1.8.8** 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

\[ x = 3t + 2, \quad y = 2t + 1, \quad z = t - 3, \quad t \in \mathbb{R}. \]

Written in terms of vectors this is

\[ (2, 1, -3) + t (3, 2, 1) = (x, y, z), \quad t \in \mathbb{R}. \]

**Example 1.8.9** A relation such as \( x^2 + y^2/4 + z^2/9 = 1 \) describes something called a level surface. It consists of the points in \( \mathbb{R}^n \), \((x, y, z)\) which satisfy the relation. Now here are parametric equations for a line: \( x = t, y = 1 + 2t, z = 1 - t \). Find where this line intersects the above level surface.

This sort of problem is not hard if you don’t panic. The points on the line are of the form \((t, 1 + 2t, 1 - t)\) where \( t \in \mathbb{R} \). All you have to do is to find values of \( t \) where this also satisfies the condition for being on the level surface. Thus you need \( t \) such that

\[ (t)^2 + (1 + 2t)^2 /4 + (1 - t)^2 /9 = 1. \]

This is just a quadratic equation. Expanding the left side yields \( \frac{19}{9}t^2 + \frac{13}{36} + \frac{7}{9}t \) and so you have to solve the quadratic equation,

\[ \frac{19}{9}t^2 + \frac{13}{36} + \frac{7}{9}t = 1 \]

First simplify this to get the equation

\[ 76t^2 + 28t - 23 = 0. \]

Then the quadratic formula gives two solutions for \( t, t = \frac{-7}{38} + \frac{9}{38}\sqrt{6}, \frac{-7}{38} - \frac{9}{38}\sqrt{6} \). Now you can obtain two points of intersection by plugging these values of \( t \) into the equation for the line. The two points are

\[ \left( -\frac{7}{38} + \frac{9}{38}\sqrt{6}, \frac{12}{19}, \frac{9}{19} \sqrt{6}, \frac{45}{38} - \frac{9}{38} \sqrt{6} \right) \]

and

\[ \left( -\frac{7}{38} - \frac{9}{38}\sqrt{6}, \frac{12}{19}, \frac{9}{19} \sqrt{6}, \frac{45}{38} + \frac{9}{38} \sqrt{6} \right). \]

Possibly you would not have guessed these points. You likely would not have found them by drawing a picture either.
1.9 Vectors And Physics

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. Also you can see that the concept of a geometric vector is useful for defining something like force.

**Definition 1.9.1** 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 1.8.5 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 points of \( \mathbb{R}^n \).

**Definition 1.9.2** Let \( \mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n \). The **position vector** of this point is the vector whose point is at \( \mathbf{x} \) and whose tail is at the origin, \( (0, \cdots, 0) \). If \( \mathbf{x} = (x_1, \cdots, x_n) \) is called a vector, the vector which is meant is this position vector just described. Another
term associated with this is **standard position.** A vector is in standard position if the tail is placed at the origin.

It is customary to identify the point in \( \mathbb{R}^n \) with its position vector. The magnitude of a vector determined by a directed line segment \( \vec{pq} \) 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 \( v \) any vector in \( \mathbb{R}^n \) the magnitude of \( v \) equals \( \left( \sum_{k=1}^{n} v_k^2 \right)^{1/2} = |v| \).

**Example 1.9.3** Consider the vector, \( v = (1, 2, 3) \) in \( \mathbb{R}^3 \). Find \( |v| \).

First, the vector is the directed line segment (arrow) which has its base at \( 0 = (0, 0, 0) \) and its point at \( (1, 2, 3) \). Therefore,

\[
|v| = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14}.
\]

What is the geometric significance of scalar multiplication? As noted earlier, if a vector, \( v \) if \( a \) represents the vector, \( v \) 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 \( a \), what is \( rv \)?

\[
|r v| = \left( \sum_{k=1}^{n} (ra_k)^2 \right)^{1/2} = \left( \sum_{k=1}^{n} r^2 (a_k)^2 \right)^{1/2} = (r^2)^{1/2} \left( \sum_{k=1}^{n} a_k^2 \right)^{1/2} = |r| |v|.
\]

Thus the magnitude of \( rv \) equals \(|r| \) times the magnitude of \( v \). If \( r \) is positive, then the vector represented by \( rv \) has the same direction as the vector, \( v \) 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 \( a \). If \( r < 0 \) similar considerations apply except in this case all the \( a_i \) also change sign. From now on, \( a \) 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.

Note there are \( n \) special vectors which point along the coordinate axes. These are

\[
e_i = (0, \cdots, 0, 1, 0, \cdots, 0)
\]

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 $e_i$ is referred to as the $i^{th}$ direction. Given a vector, $v = (a_1, \cdots, a_n)$, $a_i e_i$ is the $i^{th}$ component of the vector. Thus $a_i e_i = (0, \cdots, 0, a_i, 0, \cdots, 0)$ 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 $v = (a_1, \cdots, a_n)$,

$$v = \sum_{k=1}^{n} a_k e_i.$$ 

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 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 $a = \sum_{k=1}^{n} a_k e_i$ and $b = \sum_{k=1}^{n} b_k e_i$. Then the vector, $a$ involves a component in the $i^{th}$ direction, $a_i e_i$ while the component in the $i^{th}$ direction of $b$ is $b_i e_i$. Then it seems physically reasonable that the resultant vector should have a component in the $i^{th}$ direction equal to $(a_i + b_i) e_i$. This is exactly what is obtained when the vectors, $a$ and $b$ are added.

$$a + b = (a_1 + b_1, \cdots, a_n + b_n).$$

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 $u, v$ are vectors,

$$u = (u_1, \cdots, u_n), \quad v = (v_1, \cdots, v_n)$$

Then $u + v = (u_1 + v_1, \cdots, u_n + v_n)$. 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, $u + v$. In other words, place the vector $u$ in standard position with its base at the origin and then slide the vector $v$ till its base coincides with the point of $u$. The point of this slid vector, determines $u + v$. To illustrate, see the following picture.

Note the vector $u + v$ is the diagonal of a parallelogram determined from the two vectors $u$ and $v$ and that identifying $u + v$ with the directed diagonal of the parallelogram determined by the vectors $u$ and $v$ amounts to the same thing as the above procedure.

1.10 The Special Unit Vectors $i, j, k$

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 $i$ for $e_1$, $j$ for $e_2$, and $k$ for $e_3$. Now here are some applications of vector addition to some problems.
Example 1.10.1 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 1.10.2 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 $(100/\sqrt{2})\mathbf{i}+(100/\sqrt{2})\mathbf{j}$.

This example also motivates the concept of velocity.

Definition 1.10.3 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 $(100/\sqrt{2})\mathbf{i}+(100/\sqrt{2})\mathbf{j}$.

Example 1.10.4 The velocity of an airplane is $100\mathbf{i}+\mathbf{j}+\mathbf{k}$ 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 $\mathbf{i}$ direction a distance of $100 \times \frac{1}{60} = \frac{5}{3}$ kilometer. In the $\mathbf{j}$ direction it has moved $\frac{1}{60}$ kilometer during this same time, while it moves $\frac{1}{60}$ kilometer in the $\mathbf{k}$ 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 1.10.5 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 swam 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 $3\mathbf{j}$ while the velocity of the river would be $-4\mathbf{i}$. Therefore, the velocity of the swimmer is $-4\mathbf{i} + 3\mathbf{j}$. 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

$$\sqrt{(5/6)^2 - (1/2)^2} = \frac{2}{3} \text{ miles.}$$
Chapter 2

Vector Products

2.1 The Dot Product

Quiz

1. Given two points in \( \mathbb{R}^3 \), \((x_1, y_1, z_1)\) and \((x_2, y_2, z_2)\), show the point \( \left( \frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}, \frac{z_1 + z_2}{2} \right) \) is on the line between these two points and is the same distance from each of them.

2. Given the two points in \( \mathbb{R}^3 \), \((x_1, y_1, z_1)\) and \((x_2, y_2, z_2)\), describe the set of all points which are equidistant from these two points in terms of a simple equation.

3. An airplane heads due north at a speed of 120 miles per hour. The wind is blowing north east at a speed of 30 miles per hour. Find the resulting speed of the airplane.

2.1.1 Definition In terms Of Coordinates

There are two ways of multiplying vectors which are of great importance in applications. The first of these is called the \textbf{dot product}, also called the \textbf{scalar product} and sometimes the \textbf{inner product}.

**Definition 2.1.1** 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 2.1.2** The dot product satisfies the following properties.

\[
\begin{align*}
\mathbf{a} \cdot \mathbf{b} &= \mathbf{b} \cdot \mathbf{a} \\
\mathbf{a} \cdot \mathbf{a} &\geq 0 \text{ and equals zero if and only if } \mathbf{a} = \mathbf{0} \\
(\alpha \mathbf{a} + \beta \mathbf{b}) \cdot \mathbf{c} &= \alpha (\mathbf{a} \cdot \mathbf{c}) + \beta (\mathbf{b} \cdot \mathbf{c}) \\
\mathbf{c} \cdot (\alpha \mathbf{a} + \beta \mathbf{b}) &= \alpha (\mathbf{c} \cdot \mathbf{a}) + \beta (\mathbf{c} \cdot \mathbf{b}) \\
|\mathbf{a}|^2 &= \mathbf{a} \cdot \mathbf{a}
\end{align*}
\]
You should verify these properties. Also be sure you understand that it follows from the first three and is therefore redundant. It is listed here for the sake of convenience.

**Example 2.1.3** Find \((1, 2, 0, -1) \cdot (0, 1, 2, 3)\).

This equals \(0 + 2 + 0 - 3 = -1\).

**Example 2.1.4** Find the magnitude of \(a = (2, 1, 4, 2)\). That is, find \(|a|\).

This is \(\sqrt{(2, 1, 4, 2) \cdot (2, 1, 4, 2)} = 5\).

### 2.1.2 The Geometric Meaning Of The Dot Product, The Included Angle

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 a, b, and a-b with angle theta](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,

\[ a \cdot b = |a| |b| \cos \theta. \quad (2.6) \]

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 2.1.5** Find the angle between the vectors \(2\mathbf{i} + \mathbf{j} - \mathbf{k}\) and \(3\mathbf{i} + 4\mathbf{j} + \mathbf{k}\).

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 (2.6) 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
2.1. THE DOT PRODUCT

is \( \theta = 0.76616 \) radians or in terms of degrees, \( \theta = 0.76616 \times \frac{360}{2\pi} = 43.898^\circ \). Recall how this last computation is done. Set up a proportion: \( \frac{\pi}{2} = \frac{360}{2\pi} \) because 360° 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 2.1.6** Find the magnitude of the vector \( 2i + 3j - k \)

As discussed above, this has magnitude equal to

\[
\sqrt{(2i + 3j - k) \cdot (2i + 3j - k)} = \sqrt{4 + 9 + 1} = \sqrt{14}.
\]

**Example 2.1.7** Let \( \mathbf{u}, \mathbf{v} \) 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 \( \mathbf{u} \) and the positive \( x \) axis equals 30° and the angle between \( \mathbf{v} \) and the positive \( x \) axis is -30°. Find \( \mathbf{u} \cdot \mathbf{v} \).

From the geometric description of the dot product in 2.6

\[
\mathbf{u} \cdot \mathbf{v} = 3 \times 4 \times \cos(60^\circ) = 3 \times 4 \times 1/2 = 6.
\]

**Observation 2.1.8** Two vectors are said to be perpendicular or orthogonal if the included angle is \( \pi/2 \) radians (90°). 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 2.1.9** Determine whether the two vectors, \( 2i + j - k \) and \( \mathbf{i} + 3j + 5k \) are perpendicular.

When you take this dot product you get \( 2 + 3 - 5 = 0 \) and so these two are indeed perpendicular.

**Definition 2.1.10** When two lines intersect, the angle between the two lines is the smaller of the two angles determined.

**Example 2.1.11** 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

\[
\cos \theta = \frac{-6}{14} = \frac{-3}{7}. \quad (2.7)
\]

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.
CHAPTER 2. VECTOR PRODUCTS

2.1.3 The Cauchy Schwarz Inequality

The dot product satisfies a fundamental inequality known as the Cauchy Schwarz inequality.

**Theorem 2.1.12** The dot product satisfies the inequality

\[ |\mathbf{a} \cdot \mathbf{b}| \leq |\mathbf{a}| \ |\mathbf{b}|. \]  

(2.8)

Furthermore equality is obtained if and only if one of \( \mathbf{a} \) or \( \mathbf{b} \) is a scalar multiple of the other.

**Geometric Proof:** From the geometric description of the dot product,

\[ |\mathbf{a} \cdot \mathbf{b}| = ||\mathbf{a}|| \ |\mathbf{b}| \cos \theta \leq |\mathbf{a}| \ |\mathbf{b}| \]

because \( \cos \theta \) is a number between \(-1\) and \(1\). Equality occurs if and only if \( \cos \theta = \pm 1 \). This corresponds to \( \mathbf{b} \) being a scalar multiple of \( \mathbf{a} \). If \( \cos \theta = -1 \), then \( \mathbf{b} \) points in the opposite direction to \( \mathbf{a} \) and if \( \cos \theta = 1 \) then \( \mathbf{b} \) points in the same direction as \( \mathbf{a} \).

The Cauchy Schwarz inequality is important in many contexts other than vectors in \( \mathbb{R}^n \). What follows is a vastly superior algebraic proof. In general it is this way. Algebraic methods are nearly always to be preferred to geometric reasoning.

**Algebraic Proof:** First note that if \( \mathbf{b} = \mathbf{0} \) both sides of (2.8) 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 2.2, \( f(t) \geq 0 \) for all \( t \in \mathbb{R} \). Also from 2.3, 2.4, and 2.5

\[ f(t) = \mathbf{a} \cdot (\mathbf{a} + t\mathbf{b}) + t\mathbf{b} \cdot (\mathbf{a} + t\mathbf{b}) \]

\[ = \mathbf{a} \cdot \mathbf{a} + t(\mathbf{a} \cdot \mathbf{b}) + t^2\mathbf{b} \cdot \mathbf{b} \]

\[ = |\mathbf{a}|^2 + 2t(\mathbf{a} \cdot \mathbf{b}) + |\mathbf{b}|^2t^2. \]

Now

\[ f(t) = |\mathbf{b}|^2 \left( t^2 + 2t \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} + \frac{|\mathbf{a}|^2}{|\mathbf{b}|^2} \right) \]

\[ = |\mathbf{b}|^2 \left( t^2 + 2t \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} + \left( \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} \right)^2 - \left( \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} \right)^2 + \frac{|\mathbf{a}|^2}{|\mathbf{b}|^2} \right) \]

\[ = |\mathbf{b}|^2 \left( t + \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} \right)^2 + \left( \frac{|\mathbf{a}|^2}{|\mathbf{b}|^2} - \left( \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} \right)^2 \right) \geq 0 \]

for all \( t \in \mathbb{R} \). In particular \( f(t) \geq 0 \) when \( t = -\left( \mathbf{a} \cdot \mathbf{b} / |\mathbf{b}|^2 \right) \), the value of \( t \) which yields the minimum value of \( f \), which implies

\[ \frac{|\mathbf{a}|^2}{|\mathbf{b}|^2} = \left( \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} \right)^2 \geq 0. \]

(2.9)

Multiplying both sides by \( |\mathbf{b}|^4 \),

\[ |\mathbf{a}|^2 |\mathbf{b}|^2 \geq (\mathbf{a} \cdot \mathbf{b})^2 \]
which yields \( 2.8 \). If equality in the Cauchy Schwarz inequality holds, then the minimum value of \( f(t) \) is zero and so for some \( t, (a+tb) \cdot (a+tb) = |a+tb| = 0 \) so that \( a = -tb \). This proves the theorem.

**Another Algebraic Proof:** Let \( f(t) \) be given as above. Thus as above \( f(t) \geq 0 \) for all \( t \in \mathbb{R} \). Thus as above,

\[
f(t) = |a|^2 + 2t(a \cdot b) + |b|^2 t^2 \geq 0
\]

The graph of \( f(t) \) is a parabola which must open up and cannot cross the \( t \) axis. Thus \( f(t) = 0 \) has either one real root or no real roots. Now recall the quadratic formula. This condition implies the stuff under the square root sign in the quadratic formula must be nonpositive. Applied to this function of \( t \) it says

\[
4(a \cdot b)^2 - 4|a|^2 |b|^2 \leq 0
\]

which is just the Cauchy Schwarz inequality. As before, equality in this inequality implies \( f \) has one real zero. Thus the minimum value of \( f \) is 0. This means \( a + tb = 0 \) for some \( t \) and so one vector is a multiple of the other. This proves the theorem.

You should note that the algebraic arguments were based only on the properties of the dot product listed in \( 2.1 - 2.5 \). This means that whenever something satisfies these properties, the Cauchy Schwartz inequality holds. There are many other instances of these properties besides vectors in \( \mathbb{R}^n \).

### 2.1.4 The Triangle Inequality

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 2.1.13 (Triangle inequality)** For \( a, b \in \mathbb{R}^n \)

\[
|a + b| \leq |a| + |b| \tag{2.10}
\]

and equality holds if and only if one of the vectors is a nonnegative scalar multiple of the other. Also

\[
||a| - |b|| \leq |a - b| \tag{2.11}
\]

**Proof:** By properties of the dot product and the Cauchy Schwartz inequality,

\[
|a + b|^2 = (a + b) \cdot (a + b) \\
= (a \cdot a) + (a \cdot b) + (b \cdot a) + (b \cdot b) \\
= |a|^2 + 2(a \cdot b) + |b|^2 \\
\leq |a|^2 + 2|a \cdot b| + |b|^2 \\
\leq |a|^2 + 2|a||b| + |b|^2 \\
= (|a| + |b|)^2.
\]

Taking square roots of both sides you obtain \( 2.10 \).

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 \( 2.1.12 \) implies one of the vectors must be a multiple of
the other. Say \( \mathbf{b} = \alpha \mathbf{a} \). If \( \alpha < 0 \) then equality cannot occur in the first inequality because in this case
\[
(\mathbf{a} \cdot \mathbf{b}) = \alpha |\mathbf{a}|^2 < 0 < |\alpha| |\mathbf{a}|^2 = |\mathbf{a} \cdot \mathbf{b}|
\]
Therefore, \( \alpha \geq 0 \).

To get the other form of the triangle inequality,
\[
\mathbf{a} = \mathbf{a} - \mathbf{b} + \mathbf{b}
\]
so
\[
|\mathbf{a}| = |\mathbf{a} - \mathbf{b} + \mathbf{b}|
\leq |\mathbf{a} - \mathbf{b}| + |\mathbf{b}|.
\]
Therefore,
\[
|\mathbf{a}| - |\mathbf{b}| \leq |\mathbf{a} - \mathbf{b}|
\tag{2.12}
\]
Similarly,
\[
|\mathbf{b}| - |\mathbf{a}| \leq |\mathbf{b} - \mathbf{a}| = |\mathbf{a} - \mathbf{b}|.
\tag{2.13}
\]
It follows from (2.12) and (2.13) that (2.11) holds. This is because
\[||\mathbf{a} - \mathbf{b}|| \text{ equals the left side of either } (2.12) \text{ or } (2.13) \text{ and either way, } ||\mathbf{a} - \mathbf{b}|| \leq |\mathbf{a} - \mathbf{b}|.\]
This proves the theorem.

### 2.1.5 Direction Cosines

Now that the dot product and distance has been defined, it is possible to mention some archaic terminology which is sometimes found.

**Definition 2.1.14** Let \( \mathbf{b} \) be a vector. Then let \( \alpha \) be the angle it makes with the positive \( x \) axis, \( \beta \) the angle it makes with the positive \( y \) axis and \( \gamma \) the angle it makes with the positive \( z \) axis. The direction cosines are \( \cos(\alpha) \), \( \cos(\beta) \), and \( \cos(\gamma) \).

**Example 2.1.15** Find a formula for the direction cosines of the nonzero vector \( \mathbf{u} \).

This is really easy. From the geometric description of the dot product, the cosine of \( \alpha \) the angle between \( \mathbf{u} \) and \( \mathbf{i} \) is
\[
\cos(\alpha) = \frac{\mathbf{u} \cdot \mathbf{i}}{|\mathbf{u}||\mathbf{i}|} = \frac{\mathbf{u} \cdot \mathbf{i}}{|\mathbf{u}|}
\]
Similarly
\[
\cos(\beta) = \frac{\mathbf{u} \cdot \mathbf{j}}{|\mathbf{u}||\mathbf{j}|} = \frac{\mathbf{u} \cdot \mathbf{j}}{|\mathbf{u}|}
\]
\[
\cos(\gamma) = \frac{\mathbf{u} \cdot \mathbf{k}}{|\mathbf{u}||\mathbf{k}|} = \frac{\mathbf{u} \cdot \mathbf{k}}{|\mathbf{u}|}
\]
Note
\[
\mathbf{u} = u_1 \mathbf{i} + u_2 \mathbf{j} + u_3 \mathbf{k}
\]
where the components of \( \mathbf{u} \) are \( u_1, u_2, u_3 \). It turns out these are just the direction cosines times \( |\mathbf{u}| \). To see this, take the dot product with \( \mathbf{i} \) this yields
\[
|\mathbf{u}| \cos(\alpha) = \mathbf{u} \cdot \mathbf{i} = u_1
\]
Similarly
\[
u_2 = |\mathbf{u}| \cos(\beta), \quad u_3 = |\mathbf{u}| \cos(\gamma).
\]

**Example 2.1.16** Find the direction cosines of the vector \( (1, 2, 3) \)
\[
\cos(\alpha) = \frac{1}{\sqrt{14}}, \quad \cos(\beta) = \frac{2}{\sqrt{14}}, \quad \cos(\gamma) = \frac{3}{\sqrt{14}}.
\]
2.1.6 Work

An important application of the dot product is 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 (This is made more precise below.). 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 p₁ to p₂. There are two vectors shown, F|| and F⊥ and the picture is intended to indicate that when you add these two vectors you get F while F|| acts in the direction of motion and F⊥ acts perpendicular to the direction of motion. Only F|| contributes to the work done by F on the object as it moves from p₁ to p₂. F|| is called the projection of the force in the direction of motion. From trigonometry, you see the magnitude of F|| should equal |F||cosθ|. Thus, since F|| points in the direction of the vector from p₁ to p₂, the total work done should equal

|F|| ||p₂−p₁||cosθ = |F|| ||p₂−p₁||cosθ

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 p₁ to p₂ but in this case, cosθ 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|| ||p₂−p₁||cosθ = F·(p₂−p₁) .

This explains the following definition.

**Definition 2.1.17** Let F be a force acting on an object which moves from the point, p₁ to the point p₂. Then the work done on the object by the given force equals F·(p₂−p₁).

**Example 2.1.18** Suppose the force is F = 3i + 4j − 2k. Find the work done by this force on an object which moves in a straight line from the point (1, 2, 2) to (3, −2, 1).
From the above discussion you form the vector 
\((3, -2, 1) - (1, 2, 2) = (2, -4, -1)\).
And then you take the dot product of this vector with the given force vector.

\[
\begin{pmatrix} 2 \\ -4 \\ -1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 4 \\ -2 \end{pmatrix} = -8
\]

Thus in this case, the work impedes the motion.

**Example 2.1.19** Suppose the force is \(F = 3i + j - 2k\). Find the work done by this force on an object which moves in a straight line from the point \((1, 2, 2)\) to \((3, 6, 1)\).

From the above discussion you form the vector 
\((3, 6, 1) - (1, 2, 2) = (2, 4, -1)\).
And then you take the dot product of this vector with the given force vector.

\[
\begin{pmatrix} 2 \\ 4 \\ -1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 1 \\ -2 \end{pmatrix} = 12
\]

Thus in this case, the work aids the motion. Of course the units on work are foot pounds or Newton meters. This last is also called joules.

This is how you find work. You dot the force with the displacement vector. This is the work. It measures the extent to which the force contributes to the motion.

### 2.1.7 Projections

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 2.1.20** Let \(F\) and \(D\) be nonzero vectors. Then there exist unique vectors \(F\parallel\) and \(F\perp\) such that

\[
F = F\parallel + F\perp
\]

(2.14)

where \(F\parallel\) is a scalar multiple of \(D\), also referred to as

\[
P_D (F)
\]

which equals

\[
\frac{F \cdot D}{|D|^2} D
\]

Also \(F\perp \cdot D = 0\). The vector \(P_D (F)\) is called the projection of \(F\) onto \(D\).

**Proof:** Suppose \(\alpha D\) and \(F\parallel = \alpha D\). Taking the dot product of both sides with \(D\) and using \(F\perp \cdot D = 0\), this yields

\[
F \cdot D = \alpha |D|^2
\]

which requires \(\alpha = F \cdot D / |D|^2\). Thus there can be no more than one vector, \(F\parallel\). It follows \(F\perp\) must equal \(F - F\parallel\). This verifies there can be no more than one choice for both \(F\parallel\) and \(F\perp\).

Now let

\[
F\parallel \equiv \frac{F \cdot D}{|D|^2} D
\]
2.1. THE DOT PRODUCT

and let
\[ F_\perp = F - F_{\parallel} = F - \frac{F \cdot D}{|D|^2}D \]

Then \( F_{\parallel} = \alpha D \) where \( \alpha = \frac{F \cdot D}{|D|^2} \). It only remains to verify \( F_\perp \cdot D = 0 \). But

\[
F_\perp \cdot D = F \cdot D - \frac{F \cdot D}{|D|^2}D \cdot D
\]
\[ = F \cdot D - F \cdot D = 0. \]

This proves the theorem.

**Definition 2.1.21** The component of the vector \( F \) in the direction, \( D \) equals the scalar
\[
\frac{F \cdot D}{|D|^2}.
\]

Thus
\[
P_D(F) = \frac{F \cdot D}{|D|} \frac{D}{|D|}.
\]

In words, the projection of \( F \) on \( D \) equals the component of \( F \) in the direction \( D \) times the unit vector in the direction of \( D \).

---

The Projection Procedure

**Procedure 2.1.22** To find \( P_v(u) \), the projection of \( u \) onto \( v \) do the following.

1. Take the dot product of \( u \) with the unit vector having the same direction as \( v \). This is the component of the vector \( u \) in the direction \( v \). Thus find
\[
\frac{u \cdot v}{|v|}.
\]

2. Then you take this component and multiply by the unit vector in the direction of \( v \).
\[
P_v(u) = \frac{u \cdot v}{|v|} \frac{v}{|v|} = \frac{u \cdot v}{|v|^2} v.
\]

To write \( u \) as a sum of two vectors, one which is parallel to \( u \) and the other which is perpendicular, do the following.
\[
u = P_v(u) + (u - P_v(u))
\]

The vector \( P_v(u) \) is parallel to \( v \) and the vector \( u - P_v(u) \) is perpendicular to \( v \).

**Example 2.1.23** Let \( F = (1, 2, 3) \). Find the projection of \( F \) on \( D = (2, 1, 1) \) and also find the component of \( F \) in the direction, \( D \).
\[
P_D(F) = \frac{F \cdot D}{|D|} \frac{D}{|D|} = \frac{2 + 2 + 3}{\sqrt{4 + 1 + 1}} \frac{(2, 1, 1)}{\sqrt{4 + 1 + 1}} = \frac{7}{6} (2, 1, 1) = \left( \frac{7}{3}, \frac{7}{6}, \frac{7}{6} \right)
\]
and the component of \(F\) in the direction of \(D\) is
\[
\frac{2 + 2 + 3}{\sqrt{4 + 1 + 1}} = \frac{7}{6} \sqrt{6}.
\]

**Example 2.1.24** Let \(F = 2i + 7j - 3k\) 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
\[
(2i + 7j - 3k) \cdot (-10i - 5j + k) = -20 + (-35) + (-3) = -58 \text{ Newton meters}.
\]

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 2.1.25** Find \(P_u(v)\) if \(u = 2i + 3j - 4k\) and \(v = i - 2j + k\).

From the above discussion in Theorem 2.1.20, 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 2.1.26** Suppose \(a\), and \(b\) are vectors and \(b_\perp = b - P_a(b)\). What is the magnitude of \(b_\perp\) in terms of the included angle?

\[
|b_\perp|^2 = |b - P_a(b)\cdot(b - P_a(b))|
\]
\[
= \left( b - \frac{b \cdot a}{|a|^2} a \right) \cdot \left( b - \frac{b \cdot a}{|a|^2} a \right)
\]
\[
= |b|^2 - 2 \left( \frac{b \cdot a}{|a|^2} \right) + \left( \frac{b \cdot a}{|a|^2} \right)^2 |a|^2
\]
\[
= |b|^2 \left( 1 - \frac{(b \cdot a)^2}{|a|^2 |b|^2} \right)
\]
\[
= |b|^2 \left( 1 - \cos^2 \theta \right) = |b|^2 \sin^2 (\theta)
\]
where $\theta$ is the included angle between $\mathbf{a}$ and $\mathbf{b}$ which is less than $\pi$ radians. Therefore, taking square roots,

$$|\mathbf{b}_\perp| = |\mathbf{b}| \sin \theta.$$  

### 2.2 The Cross Product

#### Quiz

1. Find the cosine of the angle between the two vectors $\mathbf{v_1} = (1, 2, 0)$ and $\mathbf{v_2} = (2, 0, 1)$.

2. Suppose $\mathbf{u}, \mathbf{v}$ are vectors. Show the parallelogram identity.

   $$|\mathbf{u} + \mathbf{v}|^2 + |\mathbf{u} - \mathbf{v}|^2 = 2|\mathbf{u}|^2 + 2|\mathbf{v}|^2$$

   You must show this in any dimension, not just in two or three dimensions.

3. Find the projection of the vector $(1, 2, 3)$ onto the vector $(2, 3, 1)$.

4. Given two vectors, $\mathbf{u}, \mathbf{v}$ in $\mathbb{R}^n$, show using the properties of the dot product alone that

   $$\mathbf{u} - \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{v}|^2} \mathbf{v}$$

   is perpendicular to $\mathbf{v}$.

5. $\mathbf{x} = \mathbf{u} + t\mathbf{v}$ for $t \in \mathbb{R}$ is a line. Suppose $\mathbf{z}$ is a point in $\mathbb{R}^n$. Find a formula for the distance between $\mathbf{z}$ and this line.

#### 2.2.1 Cross Product, Geometric Description

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 2.2.1** Three vectors, $\mathbf{a}, \mathbf{b}, \mathbf{c}$ form a right handed system if when you extend the fingers of your right hand along the vector, $\mathbf{a}$ and close them in the direction of $\mathbf{b}$, the thumb points roughly in the direction of $\mathbf{c}$.

For an example of a right handed system of vectors, see the following picture.
In this picture the vector \( \mathbf{c} \) points upwards from the plane determined by the other two vectors. You should consider how a right hand system would differ from a left hand system. Try using your left hand and you will see that the vector, \( \mathbf{c} \) would need to point in the opposite direction as it would for a right hand system.

From now on, the vectors, \( \mathbf{i}, \mathbf{j}, \mathbf{k} \) will always form a right handed system. To repeat, if you extend the fingers of your right hand along \( \mathbf{i} \) and close them in the direction \( \mathbf{j} \), the thumb points in the direction of \( \mathbf{k} \).

The following is the geometric description of the cross product. It gives both the direction and the magnitude and therefore specifies the vector.

**Definition 2.2.2** Let \( \mathbf{a} \) and \( \mathbf{b} \) be two vectors in \( \mathbb{R}^3 \). Then \( \mathbf{a} \times \mathbf{b} \) is defined by the following two rules.

1. \( |\mathbf{a} \times \mathbf{b}| = |\mathbf{a}| |\mathbf{b}| \sin \theta \) where \( \theta \) is the included angle.
2. \( \mathbf{a} \times \mathbf{b} \cdot \mathbf{a} = 0, \mathbf{a} \times \mathbf{b} \cdot \mathbf{b} = 0, \) and \( \mathbf{a}, \mathbf{b}, \mathbf{a} \times \mathbf{b} \) forms a right hand system.

Note that \( |\mathbf{a} \times \mathbf{b}| \) is the area of the parallelogram spanned by \( \mathbf{a} \) and \( \mathbf{b} \).

The cross product satisfies the following properties.

\[ \mathbf{a} \times \mathbf{b} = - (\mathbf{b} \times \mathbf{a}) \text{ , } \mathbf{a} \times \mathbf{a} = 0, \] (2.15)

For \( \alpha \) a scalar,

\[ (\alpha \mathbf{a}) \times \mathbf{b} = \alpha (\mathbf{a} \times \mathbf{b}) = \alpha (\mathbf{a} \times \mathbf{b}) \text{ , } \] (2.16)

For \( \mathbf{a}, \mathbf{b}, \text{ and } \mathbf{c} \) vectors, one obtains the distributive laws,

\[ \mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}, \] (2.17)

\[ (\mathbf{b} + \mathbf{c}) \times \mathbf{a} = \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}. \] (2.18)

Formula (2.15) 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 (2.16) 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 \( \alpha (\mathbf{a} \times \mathbf{b}) \) and \( \mathbf{a} \times (\alpha \mathbf{b}) \). Using this yields equality in (2.16). 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 (2.16),

\[ (\mathbf{b} + \mathbf{c}) \times \mathbf{a} = -\mathbf{a} \times (\mathbf{b} + \mathbf{c}) \]
\[ = - (\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}) \]
\[ = \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}. \]

A proof of the distributive law is given in a later section for those who are interested.
2.2. THE CROSS PRODUCT

2.2.2 The Coordinate Description Of The Cross Product

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 2.2.3 Let \( a = a_1 i + a_2 j + a_3 k \) and \( b = b_1 i + b_2 j + b_3 k \) be two vectors. Then
\[
a \times b = (a_2 b_3 - a_3 b_2)i + (a_3 b_1 - a_1 b_3)j + (a_1 b_2 - a_2 b_1)k.
\] (2.19)

Proof: From the above table and the properties of the cross product listed,
\[
(a_1 i + a_2 j + a_3 k) \times (b_1 i + b_2 j + b_3 k) = \\
a_1 b_2 i \times j + a_1 b_3 i \times k + a_2 b_1 j \times i + a_2 b_3 j \times k + \\
+ a_3 b_1 k \times i + a_3 b_2 k \times j
\]
\[
= a_1 b_2 k - a_1 b_3 j - a_2 b_3 k + a_2 b_1 i + a_3 b_1 j - a_3 b_2 i
\]
\[
= (a_2 b_3 - a_3 b_2)i + (a_3 b_1 - a_1 b_3)j + (a_1 b_2 - a_2 b_1)k
\] (2.20)

This proves the proposition.

It is probably impossible for most people to remember (2.19). Fortunately, there is a somewhat easier way to remember it.

\[
a \times b = \begin{vmatrix} i & j & k \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}
\] (2.21)

where you expand the determinant along the top row. This yields
\[
(a_2 b_3 - a_3 b_2)i - (a_1 b_3 - a_3 b_1)j + (a_1 b_2 - a_2 b_1)k
\] (2.22)

which is the same as (2.20).

You will see determinants later in the course but some of you have already seen them. All you need here is how to evaluate 2 \( \times \) 2 and 3 \( \times \) 3 determinants.
\[
\begin{vmatrix} x & y \\ z & w \end{vmatrix} = xw - yz
\]

and
\[
\begin{vmatrix} a & b & c \\ x & y & z \\ u & v & w \end{vmatrix} = a \begin{vmatrix} y & z \\ v & w \end{vmatrix} - b \begin{vmatrix} x & z \\ u & w \end{vmatrix} + c \begin{vmatrix} x & y \\ u & v \end{vmatrix}.
\]

Some of you are wondering what the rule is. You look at an entry in the top row and cross out the row and column which contain that entry. If the entry is in the \( i^{th} \) column, you multiply \((-1)^{i+1}\) times the determinant of the 2 \( \times \) 2 which remains. This is the cofactor. You take the element in the top row times this cofactor and add all such up.
Example 2.2.4 Find \((i - j + 2k) \times (3i - 2j + k)\).

Use 2.21 to compute this.

\[
\begin{vmatrix}
  i & j & k \\
  1 & -1 & 2 \\
  3 & -2 & 1 \\
\end{vmatrix} = \begin{vmatrix}
  -1 & 2 \\
  -2 & 1 \\
  1 & 3 \\
\end{vmatrix} i + \begin{vmatrix}
  1 & -1 \\
  3 & -2 \\
\end{vmatrix} j + k \\
= 3i + 5j + k.
\]

Example 2.2.5 Find the area of the parallelogram determined by the vectors, \((i - j + 2k)\) and \((3i - 2j + k)\). These are the same two vectors in Example 2.2.4.

From Example 2.2.4 and the geometric description of the cross product, the area is just the norm of the vector obtained in Example 2.2.4. Thus the area is \(\sqrt{9 + 25 + 1} = \sqrt{35}\).

Example 2.2.6 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 \(\frac{1}{2} \sqrt{4 + 49 + 1} = \frac{3}{2} \sqrt{6}\).

Observation 2.2.7 In general, if you have three points (vectors) in \(\mathbb{R}^3, P, Q, R\) the area of the triangle is given by

\[
\frac{1}{2} |(Q - P) \times (R - P)|.
\]

2.2.3 A Physical Application

When a magnetic field \(M\) acts on a charged particle with charge \(s\) and if the charged particle has velocity \(v\) then the resultant force is

\[
F = s v \times M
\]

Example 2.2.8 Suppose the velocity vector of a charged particle is \((-c, c, 0)\) and \(s > 0\) and the magnetic field is \((a, 1, 1)\) where \(a\) can be controlled. Suppose you want \(F\) to have the same direction as \((c/2, c/2, c/2)\). How should \(a\) be chosen to achieve this?

You want to have \((c/2, c/2, c/2) = s (-c, c, 0) \times (a, 1, 1)\). Thus you want \(\lambda (c/2, c/2, c/2) = s (c, c, -c - ca)\)
Hence you need
\[ 2s = \lambda \]
in order to match the first two components. Then you need
\[ sc = s(-c - ca) \]
and so
\[ c = -c - ca \]
Hence \( a = -2 \).

### 2.2.4 The Box Product, Triple Product

**Definition 2.2.9** 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.

You notice the area of the base of the parallelepiped, the parallelogram determined by the vectors, \( \mathbf{a} \) and \( \mathbf{b} \) has area equal to \( |\mathbf{a} \times \mathbf{b}| \) while the altitude of the parallelepiped is \( |\mathbf{c}| \cos \theta \) where \( \theta \) is the angle shown in the picture between \( \mathbf{c} \) and \( \mathbf{a} \times \mathbf{b} \). Therefore, the volume of this parallelepiped is the area of the base times the altitude which is just
\[ |\mathbf{a} \times \mathbf{b}| |\mathbf{c}| \cos \theta = \mathbf{a} \times \mathbf{b} \cdot \mathbf{c} . \]

This expression is known as the box product and is sometimes written as \([\mathbf{a}, \mathbf{b}, \mathbf{c}]\). 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.

**Example 2.2.10** 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.

\[(i + 2j - 5k) \times (i + 3j - 6k) = \begin{vmatrix} i & j & k \\ 1 & 2 & -5 \\ 1 & 3 & -6 \end{vmatrix} = 3i + j + k\]

Now take the dot product of this vector with the third which yields

\[(3i + j + k) \cdot (3i + 2j + 3k) = 9 + 2 + 3 = 14.\]

This shows the volume of this parallelepiped is 14 cubic units.

**Observation 2.2.11** Suppose you have three vectors, \(u = (a, b, c), v = (d, e, f), \) and
\(w = (g, h, i). \) Then \(u \cdot v \times w\) is given by the following.

\[
u \cdot v \times w = (a, b, c) \cdot \begin{vmatrix} i & j & k \\ d & e & f \\ g & h & i \end{vmatrix} = a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix}.
\]

The message is that to take the box product, you can simply take the determinant of the
3 \( \times \) 3 “matrix” as described above.

**Example 2.2.12** Find the volume of the parallelepiped determined by the vectors, \((1, 2, -1), (2, 1, 5), \) and
\((-3, 1, 2).\)

As just observed, it suffices to take the absolute value of the following determinant.

\[
\begin{vmatrix} 1 & 2 & -1 \\ 2 & 1 & 5 \\ -3 & 1 & 2 \end{vmatrix} = -46
\]

Thus the volume of this parallelepiped is 46.

There is a fundamental observation which comes directly from the geometric definitions
of the cross product and the dot product.

**Lemma 2.2.13** Let \(a, b, \) and \(c\) be vectors. Then \((a \times b) \cdot c = a \cdot (b \times c).\)

**Proof:** This follows from observing that either \((a \times b) \cdot c\) and \(a \cdot (b \times c)\) both give the
volume of the parallelepiped or they both give \(-1\) times the volume.
2.2.5 A Proof Of The Distributive Law For The Cross Product

Here is a proof of the distributive law for the cross product. Let \( x \) be a vector. From the above observation,

\[
x \cdot (a \times (b + c)) = (x \times a) \cdot (b + c)
\]

\[
= (x \times a) \cdot b + (x \times a) \cdot c
\]

\[
= x \cdot a \times b + x \cdot a \times c
\]

\[
= x \cdot (a \times b + a \times c).
\]

Therefore,

\[
x \cdot [a \times (b + c) - (a \times b + a \times c)] = 0
\]

for all \( x \). In particular, this holds for \( x = a \times (b + c) - (a \times b + a \times c) \) showing that \( a \times (b + c) = a \times b + a \times c \) and this proves the distributive law for the cross product.

**Example 2.2.14** Find the volume of the parallelepiped determined by the vectors,

\((-1, 2, 3), (2, -1, 1), (3, -2, 3)\)

As just explained you only have to find the following 3 \times 3 determinants.

\[
\begin{vmatrix}
-1 & 2 & 3 \\
2 & -1 & 1 \\
3 & -2 & 3 \\
\end{vmatrix}
= -1 \begin{vmatrix}
-2 & 1 \\
3 & 3 \\
\end{vmatrix}
+ 3 \begin{vmatrix}
2 & -1 \\
3 & -2 \\
\end{vmatrix}
= -8
\]

Now volume is always nonnegative so you take the absolute value of this number. The volume of the parallelepiped is 8.

2.2.6 Torque, Moment Of A Force

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, \( F \) is applied at the end of a wrench represented by the position vector, \( R \) and the angle between these two is \( \theta \). Then the tendency to turn will be \( |R| |F| \sin \theta \), which you recognize as the magnitude of the cross product of \( R \) and \( F \). If there were just one force acting at one point whose position vector is \( R \), 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 or moment can be obtained by simply adding up the torques
associated with the different forces and positions.

Example 2.2.15 Suppose $\mathbf{R}_1 = 2\mathbf{i} - \mathbf{j} + 3\mathbf{k}$, $\mathbf{R}_2 = \mathbf{i} + 2\mathbf{j} - 6\mathbf{k}$ 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 2.2.16 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 $F_3$,
$$F_3 - F_2 = -27$$
$$F_1 - F_3 = -8$$
$$F_2 - F_1 = -8$$

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.

2.2.7 Angular Velocity

Definition 2.2.17 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 is a simple 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.
2.2. THE CROSS PRODUCT

**Example 2.2.18** A wheel rotates counter clockwise about the vector \( \mathbf{i} + \mathbf{j} + \mathbf{k} \) at 60 revolutions per minute. This means that if the thumb of your right hand were to point in the direction of \( \mathbf{i} + \mathbf{j} + \mathbf{k} \) 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 \( \frac{120\pi}{\sqrt{3}} (\mathbf{i} + \mathbf{j} + \mathbf{k}) \). Note this gives what you would expect in the case the position vector to the point is perpendicular to \( \mathbf{i} + \mathbf{j} + \mathbf{k} \) 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 2.2.19** A wheel rotates counter clockwise about the vector \( \mathbf{i} + \mathbf{j} + \mathbf{k} \) at 60 revolutions per minute exactly as in Example 2.2.18. Let \( \{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\} \) denote an orthogonal right handed system attached to the rotating wheel in which \( \mathbf{u}_3 = \frac{1}{\sqrt{3}} (\mathbf{i} + \mathbf{j} + \mathbf{k}) \). Thus \( \mathbf{u}_1 \) and \( \mathbf{u}_2 \) depend on time. Find the velocity of the point of the wheel located at the point \( 2\mathbf{u}_1 + 3\mathbf{u}_2 - \mathbf{u}_3 \).

Since \( \{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\} \) is a right handed system like \( \mathbf{i}, \mathbf{j}, \mathbf{k} \), everything applies to this system in the same way as with \( \mathbf{i}, \mathbf{j}, \mathbf{k} \). Thus the cross product is given by

\[
(a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{u}_3) \times (d\mathbf{u}_1 + e\mathbf{u}_2 + f\mathbf{u}_3) = \begin{vmatrix} u_1 & u_2 & u_3 \\ a & b & c \\ d & e & f \end{vmatrix}
\]

Therefore, in terms of the given vectors \( \mathbf{u}_i \), the angular velocity vector is \( 120\pi \mathbf{u}_3 \)

the velocity of the given point is

\[
\begin{vmatrix} u_1 & u_2 & u_3 \\ 0 & 0 & 120\pi \\ 2 & 3 & -1 \end{vmatrix} = -360\pi \mathbf{u}_1 + 240\pi \mathbf{u}_2
\]

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 \( \mathbf{u}_i \) 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

\[\sqrt{(-360\pi \mathbf{u}_1 + 240\pi \mathbf{u}_2) \cdot (-360\pi \mathbf{u}_1 + 240\pi \mathbf{u}_2)}\]

which from the properties of the dot product equals

\[\sqrt{(-360\pi)^2 + (240\pi)^2} = 120\sqrt{13}\pi\]

because the \( \mathbf{u}_i \) are given to be orthogonal.
2.2.8 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 2.2.20** Let an object consist of $p$ point masses, $m_1, \cdots, m_p$ with the position of the $k$th of these at $R_k$. The center of mass of this object, $R_0$ is the point satisfying

$$
\sum_{k=1}^{p} (R_k - R_0) \times g m_k u = 0
$$

for all unit vectors, $u$.

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,

$$
\left( \sum_{k=1}^{p} R_k g m_k - R_0 \sum_{k=1}^{p} g m_k \right) \times u = 0
$$

for any choice of unit vector, $u$. You should verify that if $a \times u = 0$ for all $u$, then it must be the case that $a = 0$. Then the above formula requires that

$$
\sum_{k=1}^{p} R_k g m_k - R_0 \sum_{k=1}^{p} g m_k = 0.
$$

dividing by $g$, and then by $\sum_{k=1}^{p} m_k$,

$$
R_0 = \frac{\sum_{k=1}^{p} R_k m_k}{\sum_{k=1}^{p} m_k}.
$$

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 2.2.21** Let $m_1 = 5, m_2 = 6$, and $m_3 = 3$ where the masses are in kilograms. Suppose $m_1$ is located at $2i + 3j + k$, $m_2$ is located at $i - 3j + 2k$ and $m_3$ is located at $2i - j + 3k$. Find the center of mass of these three masses.

Using (2.2.20)

$$
R_0 = \frac{5(2i + 3j + k) + 6(i - 3j + 2k) + 3(2i - j + 3k)}{5 + 6 + 3}
$$

$$
= \frac{11i}{7} - \frac{3j}{7} + \frac{13k}{7}
$$
2.3 Further Explanations

2.3.1 The Distributive Law For The Cross Product

This section gives a proof for which is independent of volume considerations. It is included here for the interested student. If you are satisfied with taking the distributive law on faith or are happy with the other argument given, it is not necessary to read this section. The proof given here is quite clever and follows the one given in . The other approach based on areas is found in and is discussed briefly earlier.

Lemma 2.3.1 Let $b$ and $c$ be two vectors. Then $b \times c = b \times c_\perp$ where $c_\parallel + c_\perp = c$ and $c_\perp \cdot b = 0$.

Proof: Consider the following picture.

Now $c_\perp = c - c \cdot \frac{b}{|b|} \frac{b}{|b|}$ 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 2.3.2 Let $a, b$, and $c$ be vectors in $\mathbb{R}^3$. Then

$$a \times (b + c) = a \times b + a \times c$$  \hspace{1cm} (2.25)

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.

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 $a \times b$ you move counterclockwise through an angle of $\pi/2$ radians from the vector, $b$. Similar relationships exist between the vectors $a \times (b + c)$ and $b + c$ and the vectors $a \times c$ and $c$. Thus the angle between $a \times b$ and $a \times (b + c)$ is the same as the angle between $b + c$ and $b$ and the angle between $a \times c$ and $a \times (b + c)$ is the same as the angle between $c$ and $b + c$. In addition to this, since $a$ is perpendicular to these vectors,

$$|a \times b| = |a| |b|, |a \times (b + c)| = |a| |b + c|,$$

and

$$|a \times c| = |a| |c|.$$

Therefore,

$$\frac{|a \times (b + c)|}{|b + c|} = \frac{|a \times c|}{|c|} = \frac{|a \times b|}{|b|} = \frac{|a|}{|b|},$$

and so

$$\frac{|a \times (b + c)|}{|a \times c|} = \frac{|b + c|}{|c|}, \quad \frac{|a \times (b + c)|}{|a \times b|} = \frac{|b + c|}{|b|}$$

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 $\triangle$.

Now suppose it is not necessarily the case that $a \cdot b = a \cdot c = 0$. Then write $b = b_\parallel + b_\perp$ where $b_\perp \cdot a = 0$. Similarly $c = c_\parallel + c_\perp$. By the above lemma and what was just shown,

$$a \times (b + c) = a \times (b + c)_\perp$$

$$= a \times (b_\perp + c_\perp)$$

$$= a \times b_\perp + a \times c_\perp$$

$$= a \times b + a \times c.$$

This proves the theorem.

### 2.3.2 Vector Identities And Notation*

To begin with consider $u \times (v \times w)$ and it is desired to simplify this quantity. It turns out this is an important quantity which comes up in many different contexts. Let $u = (u_1, u_2, u_3)$ and let $v$ and $w$ be defined similarly.

The cross product of two vectors $v = (v_1, v_2, v_3)$ and $w = (w_1, w_2, w_3)$ is given by

$$v \times w = \begin{vmatrix} i & j & k \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}$$

$$= (v_2 w_3 - v_3 w_2)i + (w_1 v_3 - v_1 w_3)j + (v_1 w_2 - v_2 w_1)k$$

Next consider $u \times (v \times w)$ which is given by

$$u \times (v \times w) = \begin{vmatrix} i & j & k \\ u_1 & u_2 & u_3 \\ (v_2 w_3 - v_3 w_2) & (w_1 v_3 - v_1 w_3) & (v_1 w_2 - v_2 w_1) \end{vmatrix}.$$

When you multiply this out, you get

$$i \left( (v_1 u_2 w_2 + u_3 v_1 w_3 - u_1 w_2 v_2 - u_3 w_1 v_3) + j (v_2 u_1 w_1 + v_2 w_3 u_3 - w_2 u_1 v_1 - u_3 w_2 v_3) + k (u_1 v_3 w_3 + v_3 w_2 u_2 - u_1 v_3 w_3 - v_2 w_3 u_2) \right).$$
and if you are clever, you see right away that
\[(iv_1 + jv_2 + kv_3)(u_1w_1 + u_2w_2 + u_3w_3) - (iw_1 + jw_2 + kw_3)(u_1v_1 + u_2v_2 + u_3v_3).\]
Thus
\[
u \times (v \times w) = v (u \cdot w) - w (u \cdot v).
\]
A related formula is
\[
(u \times v) \times w = -[w \times (u \times v)]
\]
\[
= -[u (w \cdot v) - v (w \cdot u)]
\]
\[
= v (w \cdot u) - u (w \cdot v).
\]
This derivation is simply wretched and it does nothing for other identities which may arise in applications. Actually, the above two formulas, and 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, and which are very useful in dealing with vector identities. To begin with, here is the definition of these symbols.

**Definition 2.3.3** The symbol, , called the Kroneker delta symbol is defined as follows.
\[
\delta_{ij} = \begin{cases} 
1 & \text{if } i = j \\
0 & \text{if } i \neq j
\end{cases}
\]
With the Kroneker symbol, and can equal any integer in \(\{1, 2, \ldots, n\}\) for any \(n \in \mathbb{N}\).

**Definition 2.3.4** For , and integers in the set, \(\{1, 2, 3\}\), is defined as follows.
\[
\varepsilon_{ijk} = \begin{cases} 
1 & \text{if } (i,j,k) = (1,2,3), (2,3,1), \text{ or } (3,1,2) \\
-1 & \text{if } (i,j,k) = (2,1,3), (1,3,2), \text{ or } (3,2,1) \\
0 & \text{if there are any repeated integers}
\end{cases}
\]
The subscripts and in the above are called indices. A single one is called an index. This symbol, , is also called the permutation symbol.

The way to think of is that \(\varepsilon_{123} = 1\) and if you switch any two of the numbers in the list , 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 \(\varepsilon_{iii} = -\varepsilon_{iij}\) and so \(\varepsilon_{ij} = 0\).

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 \(a_ib_i\) means \(\sum a_ib_i\). 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 \(a_ib_i\) is all right but \(a_isb_i\) is not. The reason for this is that you end up getting confused about what is meant. If you want to write \(\sum c_i\), it is best to simply use the summation notation. There is a very important reduction identity connecting these two symbols.

**Lemma 2.3.5** 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 \( i = s \) and \( j = r \), there is exactly one term in the sum on the left which is nonzero and it must equal -1. The right side 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 2.3.6** Let \( \mathbf{u}, \mathbf{v} \) be vectors in \( \mathbb{R}^n \) where the Cartesian coordinates of \( \mathbf{u} \) are \((u_1, \cdots, u_n)\) and the Cartesian coordinates of \( \mathbf{v} \) are \((v_1, \cdots, v_n)\). Then \( \mathbf{u} \cdot \mathbf{v} = u_iv_i \). If \( \mathbf{u}, \mathbf{v} \) are vectors in \( \mathbb{R}^3 \), then

\[
(\mathbf{u} \times \mathbf{v})^i = \varepsilon_{ijk}u_jv_k.
\]

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,

\[
\mathbf{u} \times \mathbf{v} \equiv \begin{vmatrix} i & j & k \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}
\]

and so

\[
(\mathbf{u} \times \mathbf{v})^1 = (u_2v_3 - u_3v_2).
\]

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 \((\mathbf{u} \times \mathbf{v})^2\) and \((\mathbf{u} \times \mathbf{v})^3\) 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 2.3.7** Discover a formula which simplifies \((\mathbf{u} \times \mathbf{v}) \times \mathbf{w}\).

From the above reduction formula,

\[
((\mathbf{u} \times \mathbf{v}) \times \mathbf{w})^i = \varepsilon_{ijk}(\mathbf{u} \times \mathbf{v})^jw_k = \varepsilon_{ijk} \varepsilon_{jrs}u_rv_sw_k = -\varepsilon_{jik} \varepsilon_{jrs}u_rw_sw_k = -(\delta_{ir} \delta_{ks} - \delta_{is} \delta_{kr})u_rv_kw_k = -u_iv_kw_k - u_kv_iw_k = \mathbf{u} \cdot \mathbf{w}v_i - \mathbf{v} \cdot \mathbf{w}u_i = ((\mathbf{u} \cdot \mathbf{w}) \mathbf{v} - (\mathbf{v} \cdot \mathbf{w}) \mathbf{u})^i.
\]

Since this holds for all \( i \), it follows that

\[
(\mathbf{u} \times \mathbf{v}) \times \mathbf{w} = (\mathbf{u} \cdot \mathbf{w}) \mathbf{v} - (\mathbf{v} \cdot \mathbf{w}) \mathbf{u}.
\]
2.4 Exercises With Answers

1. Draw the vector $\mathbf{u} = (1, -2)$, the vector $\mathbf{v} = (2, 3)$, and the vector $(1, -2) + (2, 3) = \mathbf{u} + \mathbf{v}$.

![Diagram showing vectors u, v, and u + v]

2. Let $\mathbf{u} = (1, 2, -5)$, $\mathbf{v} = (3, -1, 2)$ and $\mathbf{w} = (2, 0, 3)$ Find the following.

(a) $(2\mathbf{u} + \mathbf{v}) \cdot \mathbf{w}$

This is $(2(1, 2, -5) + (3, -1, 2)) \cdot (2, 0, 3) = -14$. Here is why.

$2(1, 2, -5) + (3, -1, 2) = (5, 3, -8)$

and

$(5, 3, -8) \cdot (2, 0, 3) = -14$

(b) $(\mathbf{u} - 3\mathbf{v}) \cdot \mathbf{w}$

This is $((1, 2, -5) - 3(3, -1, 2)) \cdot (2, 0, 3) = -49$

3. Find the cosine of the angle between the two vectors, $(1, 2, 5)$ and $(3, -2, 1)$.

$\cos \theta = \frac{(1, 2, 5) \cdot (3, -2, 1)}{\sqrt{30} \sqrt{14}} = \frac{1}{105} \sqrt{30} \sqrt{14} = 0.19518$

4. Here are two vectors, $(1, 2, 3)$ and $(3, 2, 1)$. Find a vector which is perpendicular to both of these vectors.

One way to do this is to take the cross product of the two vectors. $(1, 2, 3) \times (3, 2, 1) = (-4, 8, -4)$. A vector perpendicular to both of these vectors is $(-1, 2, 1)$. Note nothing is changed as far as being perpendicular is concerned by division by 4.

5. Given two vectors in $\mathbb{R}^n$, $\mathbf{u}, \mathbf{v}$ show that

$\mathbf{u} \cdot \mathbf{v} = \frac{1}{4} \left( |\mathbf{u} + \mathbf{v}|^2 - |\mathbf{u} - \mathbf{v}|^2 \right)$.

This is really easy if you remember the axioms for the dot product. Otherwise it is very troublesome. Start with the right side.

$\frac{1}{4} \left( |\mathbf{u} + \mathbf{v}|^2 - |\mathbf{u} - \mathbf{v}|^2 \right) = \frac{1}{4} ((\mathbf{u} + \mathbf{v}) \cdot (\mathbf{u} + \mathbf{v}) - (\mathbf{u} - \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v}))$

$= \frac{1}{4} [\mathbf{u} \cdot \mathbf{u} + \mathbf{v} \cdot \mathbf{v} - \{\mathbf{u} \cdot \mathbf{u} + \mathbf{v} \cdot \mathbf{v} - 2\mathbf{u} \cdot \mathbf{v}\}]$

$= \frac{1}{4} [2\mathbf{u} \cdot \mathbf{v} - (\mathbf{u} \cdot \mathbf{v})] = \mathbf{u} \cdot \mathbf{v}$. 

6. If $|\mathbf{u}| = 3$, $|\mathbf{v}| = 4$, and $\mathbf{u} \cdot \mathbf{v} = 5$, find $|\mathbf{u} + \mathbf{v}|$.

This is easy if you know the properties of the dot product. Otherwise it is trouble.

$$|\mathbf{u} + \mathbf{v}|^2 = |\mathbf{u}|^2 + |\mathbf{v}|^2 + 2\mathbf{u} \cdot \mathbf{v}$$

$$= 9 + 16 + 50 = 75.$$ Therefore, $|\mathbf{u} + \mathbf{v}| = 5\sqrt{5}$.

7. Find vectors in $\mathbb{R}^3$, $\mathbf{u}, \mathbf{v}$ such that $\mathbf{u} \cdot \mathbf{v} = 6$ and $|\mathbf{u}| = 2$ while $|\mathbf{v}| = 3$. You see that equality holds in the Cauchy Schwarz inequality and so one of these vectors must be a multiple of the other. It must be a positive multiple of the other because the dot product is positive which implies the angle between the vectors is 0 and not $\pi$.

Let $\mathbf{u} = (0, 0, 2), \mathbf{v} = (0, 0, 3)$ . This appears to work. You should find some other examples. What if $\mathbf{u} = (\sqrt{2}/2, \sqrt{2}/2, \sqrt{3})$. In this case $|\mathbf{u}| = 2$ also. Can you find $\mathbf{v}$ such that the above will hold?

8. The projection of $\mathbf{u}$ onto $\mathbf{v}$, denoted by $\text{proj}_\mathbf{v}(\mathbf{u})$ is given by the formula

$$\text{proj}_\mathbf{v}(\mathbf{u}) = \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{v}|^2} \mathbf{v}$$

Show $\mathbf{u} - \text{proj}_\mathbf{v}(\mathbf{u})$ is perpendicular to $\mathbf{v}$. Also show $\text{proj}_\mathbf{v}(a\mathbf{u} + b\mathbf{w}) = a(\text{proj}_\mathbf{v}(\mathbf{u})) + b(\text{proj}_\mathbf{v}(\mathbf{w}))$.

This is another of those things which is very easy if you know the properties of the dot product but lots of trouble if you don’t. Of course you can persist in not learning these things if you want. It is up to you.

$$\mathbf{v} \cdot (\mathbf{u} - \text{proj}_\mathbf{v}(\mathbf{u})) = \mathbf{v} \cdot \left(\mathbf{u} - \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{v}|^2} \mathbf{v}\right) = \mathbf{v} \cdot \mathbf{u} - \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{v}|^2} \mathbf{v} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u} - \mathbf{u} \cdot \mathbf{v} = 0$$

This does it. The dot product equals zero and so the two vectors are perpendicular.

As to the other claim,

$$\frac{(a\mathbf{u} + b\mathbf{w}) \cdot \mathbf{v}}{|\mathbf{v}|^2} = a \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{v}|^2} + b \frac{\mathbf{w} \cdot \mathbf{v}}{|\mathbf{v}|^2} = \equiv a \left(\text{proj}_\mathbf{v}(\mathbf{u})\right) + b \left(\text{proj}_\mathbf{v}(\mathbf{w})\right).$$

Now that was real easy wasn’t it. Note I never said anything about $\mathbf{u}, \mathbf{v}$ being lists of numbers. I just used the properties of the dot product.

9. Find the angle between the vectors $3\mathbf{i} - \mathbf{j} - \mathbf{k}$ and $\mathbf{i} + 4\mathbf{j} + 2\mathbf{k}$.

$$\cos \theta = \frac{3 - 1 - 2}{\sqrt{1 + 16 + 4} \sqrt{1 + 16 + 4}} = - .19739.$$ Therefore, you have to solve the equation $\cos \theta = - .19739$. Solution is : $\theta = 1.7695$ radians. You need to use a calculator or table to solve this.

10. Find $\mathbf{P}_\mathbf{u}(\mathbf{v})$ where $\mathbf{v} = (1, 3, -2)$ and $\mathbf{u} = (1, 2, 3)$.

Remember to find this you take $\frac{\mathbf{v} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}}\mathbf{u}$. Thus the answer is $\frac{1}{11} (1, 2, 3)$. 

11. If $\mathbf{F}$ is a force and $\mathbf{D}$ is a vector, show $\mathbf{P}_\mathbf{D}(\mathbf{F}) = |\mathbf{F}| \cos \theta \mathbf{u}$ 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, $\mathbf{D}$.

$$\mathbf{P}_\mathbf{D}(\mathbf{F}) = \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2} \mathbf{D} = |\mathbf{F}| |\mathbf{D}| \cos \theta \frac{1}{|\mathbf{D}|^2} \mathbf{D} = |\mathbf{F}| \cos \theta \frac{\mathbf{D}}{|\mathbf{D}|}.$$
12. A boy drags a sled for 100 feet along the ground by pulling on a rope which is 40 degrees from the horizontal with a force of 10 pounds. How much work does this force do?

The component of force is \(10 \cos \left(\frac{40 \pi}{180}\right)\) and it acts for 100 feet so the work done is

\[10 \cos \left(\frac{40 \pi}{180}\right) \times 100 = 766.04\]

13. If \(\mathbf{a}, \mathbf{b},\) and \(\mathbf{c}\) are vectors. Show that \((\mathbf{b} + \mathbf{c})_\perp = \mathbf{b}_\perp + \mathbf{c}_\perp\) where \(\mathbf{b}_\perp = \mathbf{b} - \mathbf{P}_\mathbf{a}(\mathbf{b})\).

14. Find \((1, 0, 3, 4) \cdot (2, 7, 1, 3)\).

\[(1, 0, 3, 4) \cdot (2, 7, 1, 3) = 17\]

15. Prove from the axioms of the dot product the parallelogram identity, \(|a + b|^2 + |a - b|^2 = 2|a|^2 + 2|b|^2\).

Use the properties of the dot product and the definition of the norm in terms of the dot product. Thus the left side is

\[
\mathbf{a} \cdot \mathbf{a} + \mathbf{b} \cdot \mathbf{b} + 2(\mathbf{a} \cdot \mathbf{b}) + \mathbf{a} \cdot \mathbf{a} + \mathbf{b} \cdot \mathbf{b} - 2\mathbf{a} \cdot \mathbf{b} = 2|\mathbf{a}|^2 + 2|\mathbf{b}|^2.
\]

16. Find all vectors, \((x, y)\) which are perpendicular to \((1, 2)\).

You need \(x + 2y = 0\) so \(x = -2y\) and you can write all desired vectors in the form

\((-2y, y) : y \in \mathbb{R}\).

17. Find the line through \((1, 2, 1)\) and \((2, 0, 3)\).

First get a direction vector which in this case is \((1, -2, 2)\). Then the equation of the line is

\[(x, y, z) = (1, 2, 1) + t(1, -2, 2) = (1 + t, 2 - 2t, 1 + 2t)\]

Thus a parametric form for this line is \(x = 1 + t, y = 2 - 2t, z = 1 + 2t\) and a vector equation for this line is

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

if you want to write the vectors as column vectors.

18. In \(\mathbb{R}^2\), the equation of a line is given as \(2x + 3y = 6\). Find a vector equation of this line.

One way to do it is to get a couple of points on the line and then do as in the previous problem. Two points on this line are \((0, 2)\) and \((3, 0)\). Then a direction vector for the line is \((-3, 2)\) and so a vector equation of the line is

\[(x, y) = (0, 2) + t(-3, 2)\]

Written parametrically, this would be \(x = -3t, y = 2 + 2t\).

19. Suppose you have the vector equation for a line joining the two points, \(\mathbf{p}, \mathbf{q}\). This is

\[
\mathbf{x} = \mathbf{p} + t(\mathbf{q} - \mathbf{p})
\]
Note this works because when \( t = 0 \) the right side is \( p \) and when \( t = 1 \), the right side is \( q \). Now find the point which is \( 1/3 \) of the way between \( p \) and \( q \). Thus the point is

\[
x_{1/3} = \frac{2}{3}p + \frac{1}{3}q.
\]

Does it work?

\[
\|x_{1/3} - p\| = \left| -\frac{1}{3}p + \frac{1}{3}q \right| = \frac{1}{3}|q - p|.
\]

Seems to work just fine. I suppose you could also find points which are \( 1/5 \) of the way between \( p \) and \( q \) also.

20. The wind blows from West to East at a speed of 30 kilometers per hour and an airplane which travels at 300 Kilometers per hour in still air is heading North West. What is the velocity of the airplane relative to the ground? What is the component of this velocity in the direction North?

Let the positive \( y \) axis point in the direction North and let the positive \( x \) axis point in the direction East. The velocity of the wind is \( 30i \). The plane moves in the direction \( i + j \). A unit vector in this direction is \( \frac{1}{\sqrt{2}}(i + j) \). Therefore, the velocity of the plane relative to the ground is \( 30i + \frac{300}{\sqrt{2}}(i + j) = 150\sqrt{2}j + (30 + 150\sqrt{2})i \). The component of velocity in the direction North is \( 150\sqrt{2} \).

21. In the situation of Problem 20 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 relative to the ground?

In this case the unit vector will be \(-\sin(\theta) i + \cos(\theta) j\). Therefore, the velocity of the plane will be

\[
300(-\sin(\theta) i + \cos(\theta) j)
\]

and this is supposed to satisfy

\[
300(-\sin(\theta) i + \cos(\theta) j) + 30i = 0i + ?j.
\]

Therefore, you need to have \( \sin(\theta) = 1/10 \), which means \( \theta = 0.10017 \) radians. Therefore, the degrees should be \( \frac{1 \times 180}{\pi} = 5.7296 \) degrees. In this case the velocity vector of the plane relative to the ground is \( 300 \left( \frac{\sqrt{99}}{10} \right) j \).

22. In the situation of 21 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?

The airplane needs to fly 600 miles at a speed of \( 300 \left( \frac{\sqrt{99}}{10} \right) \). Therefore, it takes

\[
\frac{600}{(300(\frac{\sqrt{99}}{10}))} = 2.0101 \text{ hours to get there. Therefore, the plane will need to use about 68 gallons of gas. It won’t make it.}
\]

23. A certain river is one half mile wide with a current flowing at 3 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 2 miles per hour. How far down the river does he find himself when he has swam across? How far does he end up swimming?

The velocity of the man relative to the earth is then \(-3i + 2j\). Since the component of \( j \) equals 2 it follows he takes \( 1/8 \) of an hour to get across. During this time he
2.4. EXERCISES WITH ANSWERS

is swept downstream at the rate of 3 miles per hour and so he ends up 3/8 of a mile down stream. He has gone \( \sqrt{\left(\frac{3}{8}\right)^2 + \left(\frac{1}{2}\right)^2} = .625 \) miles in all.

24. Three forces are applied to a point which does not move. Two of the forces are \( 2\mathbf{i} - \mathbf{j} + 3\mathbf{k} \) Newtons and \( \mathbf{i} - 3\mathbf{j} - 2\mathbf{k} \) Newtons. Find the third force.

   Call it \( a\mathbf{i} + b\mathbf{j} + c\mathbf{k} \) Then you need \( a + 2 + 1 = 0, b - 1 - 3 = 0, \text{ and } c + 3 - 2 = 0. \)

   Therefore, the force is \(-3\mathbf{i} + 4\mathbf{j} - \mathbf{k} \).

25. If you only assume (4) holds for \( \mathbf{u} = \mathbf{i}, \mathbf{j}, \mathbf{k} \), show that this implies (2) holds for all unit vectors, \( \mathbf{u} \).

   Suppose that \( (\sum_{k=1}^{p} \mathbf{R}_k gm_k - \mathbf{R}_0 \sum_{k=1}^{p} gm_k) \times \mathbf{u} = \mathbf{0} \) for \( \mathbf{u} = \mathbf{i}, \mathbf{j}, \mathbf{k} \). Then if \( \mathbf{u} \) is an arbitrary unit vector, \( \mathbf{u} \) must be of the form \( a\mathbf{i} + b\mathbf{j} + c\mathbf{k} \). Now from the distributive property of the cross product and letting \( \mathbf{w} = (\sum_{k=1}^{p} \mathbf{R}_k gm_k - \mathbf{R}_0 \sum_{k=1}^{p} gm_k) \), this says

   \[
   (\sum_{k=1}^{p} \mathbf{R}_k gm_k - \mathbf{R}_0 \sum_{k=1}^{p} gm_k) \times \mathbf{u} = \mathbf{w} \times (a\mathbf{i} + b\mathbf{j} + c\mathbf{k}) \\
   = aw \times \mathbf{i} + bw \times \mathbf{j} + cw \times \mathbf{k} \\
   = 0 + 0 + 0 = 0.
   \]

26. Let \( m_1 = 4, m_2 = 3 \), and \( m_3 = 1 \) where the masses are in kilograms and the distance is in meters. Suppose \( m_1 \) is located at \( 2\mathbf{i} - \mathbf{j} + \mathbf{k} \), \( m_2 \) is located at \( 2\mathbf{i} - 3\mathbf{j} + \mathbf{k} \) and \( m_3 \) is located at \( 2\mathbf{i} + \mathbf{j} + 3\mathbf{k} \). Find the center of mass of these three masses.

   Let the center of mass be located at \( a\mathbf{i} + b\mathbf{j} + c\mathbf{k} \). Then \( (4 + 3 + 1) (a\mathbf{i} + b\mathbf{j} + c\mathbf{k}) = 4(2\mathbf{i} - \mathbf{j} + \mathbf{k}) + 3(2\mathbf{i} - 3\mathbf{j} + \mathbf{k}) + 1(2\mathbf{i} + \mathbf{j} + 3\mathbf{k}) = 16\mathbf{i} - 12\mathbf{j} + 10\mathbf{k} \). Therefore, \( a = 2, b = \frac{-7}{2} \) and \( c = \frac{5}{2} \). The center of mass is then \( 2\mathbf{i} - \frac{7}{2}\mathbf{j} + \frac{5}{2}\mathbf{k} \).

27. Find the angular velocity vector of a rigid body which rotates counter clockwise about the vector \( \mathbf{i} - \mathbf{j} + \mathbf{k} \) at 20 revolutions per minute. Assume distance is measured in meters.

   The angular velocity is \( 20 \times 2\pi = 40\pi \). Then \( \Omega = 40\pi \frac{1}{\sqrt{3}} (\mathbf{i} - \mathbf{j} + \mathbf{k}) \).

28. Find the area of the triangle determined by the three points, \( (1,2,3), (1,2,6) \) and \( (-3,2,1) \).

   The three points determine two displacement vectors from the point \( (1,2,3), (0,0,3) \) and \( (-4,0,-2) \). To find the area of the parallelogram determined by these two displacement vectors, you simply take the norm of their cross product. To find the area of the triangle, you take one half of that. Thus the area is

   \[
   \frac{1}{2} |(0,0,3) \times (-4,0,-2)| = \frac{1}{2} |(0,-12,0)| = 6.
   \]

29. Find the area of the parallelogram determined by the vectors, \( (1,0,3) \) and \( (4,-2,1) \).

   \[
   |(1,0,3) \times (4,-2,1)| = |(6,11,-2)| = \sqrt{36 + 121 + 4} = \sqrt{151}.
   \]

30. Find the volume of the parallelepiped determined by the vectors, \( \mathbf{i} - 7\mathbf{j} - 5\mathbf{k}, \mathbf{i} + 2\mathbf{j} - 6\mathbf{k}, 3\mathbf{i} - 3\mathbf{j} + \mathbf{k} \).

   Remember you just need to take the absolute value of the determinant having the given vectors as rows. Thus the volume is the absolute value of

   \[
   \begin{vmatrix}
   1 & -7 & -5 \\
   1 & 2 & -6 \\
   3 & -3 & 1 \\
   \end{vmatrix} = 162
   \]
31. Suppose \( \mathbf{a}, \mathbf{b}, \) and \( \mathbf{c} \) 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?

**Hint:** Consider what happens when you take the determinant of a matrix which has all integers.

32. 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. It is best if you use the geometric reasoning. Here is a picture which might help.

In this picture there is an angle between \( \mathbf{a} \times \mathbf{b} \) and \( \mathbf{c} \). Call it \( \theta \). Now if you take \( |\mathbf{a} \times \mathbf{b}| |\mathbf{c}| \cos \theta \) this gives the area of the base of the parallelepiped determined by \( \mathbf{a} \) and \( \mathbf{b} \) times the altitude of the parallelepiped, \( |\mathbf{c}| \cos \theta \). This is what is meant by the volume of the parallelepiped. It also equals \( \mathbf{a} \times \mathbf{b} \cdot \mathbf{c} \) by the geometric description of the dot product. Similarly, there is an angle between \( \mathbf{b} \times \mathbf{c} \) and \( \mathbf{a} \). Call it \( \alpha \). Then if you take \( |\mathbf{b} \times \mathbf{c}| |\mathbf{a}| \cos \alpha \) this would equal the area of the face determined by the vectors \( \mathbf{b} \) and \( \mathbf{c} \) times the altitude measured from this face, \( |\mathbf{a}| \cos \alpha \). Thus this also is the volume of the parallelepiped. and it equals \( \mathbf{a} \cdot \mathbf{b} \times \mathbf{c} \). The picture is not completely representative. If you switch the labels of two of these vectors, say \( \mathbf{b} \) and \( \mathbf{c} \), explain why it is still the case that \( \mathbf{a} \cdot \mathbf{b} \times \mathbf{c} = \mathbf{a} \times \mathbf{b} \cdot \mathbf{c} \). You should draw a similar picture and explain why in this case you get \(-1\) times the volume of the parallelepiped.

33. Discover a vector identity for \((\mathbf{u} \times \mathbf{v}) \times \mathbf{w}\).

\[
\left( (\mathbf{u} \times \mathbf{v}) \times \mathbf{w} \right)_i = \varepsilon_{ijk} (\mathbf{u} \times \mathbf{v})_j w_k = \varepsilon_{ijk} \varepsilon_{jrs} u_r v_s w_k = (\delta_{ls} \delta_{kr} - \delta_{lr} \delta_{ks}) u_r v_s w_k = u_k w_j v_i - u_i w_k v_j = (\mathbf{u} \cdot \mathbf{w}) v_i - (\mathbf{v} \cdot \mathbf{w}) u_i.
\]

Therefore, \((\mathbf{u} \times \mathbf{v}) \times \mathbf{w} = (\mathbf{u} \cdot \mathbf{w}) \mathbf{v} - (\mathbf{v} \cdot \mathbf{w}) \mathbf{u}\).

34. Discover a vector identity for \((\mathbf{u} \times \mathbf{v}) \cdot (\mathbf{z} \times \mathbf{w})\).
2.4. EXERCISES WITH ANSWERS

Start with \( \varepsilon_{ijk} u_j \varepsilon_{iws} z_w \) and then go to work on it using the reduction identities for the permutation symbol.

35. Discover a vector identity for \((u \times v) \times (z \times w)\) in terms of box products.

You will save time if you use the identity for \((u \times v) \times w\) or \(u \times (v \times w)\).
Part III

Planes And Systems Of Equations
Chapter 3

Planes

3.1 Finding Planes

Quiz

1. Let \( \mathbf{a} = (1, 2, 3), \mathbf{b} = (2, -1, 1) \). Find a vector which is perpendicular to both of these vectors.

2. Find the area of the parallelogram determined by the above two vectors.

3. Find the cosine of the angle between the above two vectors.

4. Find the sine of the angle between the above two vectors.

5. Find the volume of the parallelepiped determined by the vectors, \( \mathbf{a} = (1, 2, 3), \mathbf{b} = (2, -1, 1) \) and \( \mathbf{c} = (1, 1, 1) \).

3.1.1 The Cartesian Equation Of A Plane

A plane is a long flat thing. 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 3.1.1** 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
\[
(1, 2, 3) \cdot (x - 2, y + 1, z - 5) = x - 15 + 2y + 3z = 0
\]

**Example 3.1.2** \( 2x + 4y - 5z = 11 \) is the equation of a plane. Find the normal vector and a point on this plane.

You can write this in the form \( 2 \left( x - \frac{11}{2} \right) + 4 (y - 0) + (-5) (z - 0) = 0 \). Therefore, a normal vector to the plane is \( 2\mathbf{i} + 4\mathbf{j} - 5\mathbf{k} \) and a point in this plane is \( \left( \frac{11}{2}, 0, 0 \right) \). Of course there are many other points in the plane.

**Proposition 3.1.3** If \( (a, b, c) \neq (0, 0, 0) \), then \( ax + by + cz = d \) is the equation of a plane with normal vector \( a\mathbf{i} + b\mathbf{j} + c\mathbf{k} \). 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 \( a\mathbf{i} + b\mathbf{j} + c\mathbf{k} \). Suppose \( a \neq 0 \). Then the points which satisfy \( ax + by + cz = d \) are the same as the points which satisfy
\[
a \left( x - \frac{d}{a} \right) + b (y - 0) + c (z - 0) = 0.
\]
Thus a point on the plane is \( \left( \frac{d}{a}, 0, 0 \right) \) and a normal vector is \( (a, b, c) \) as claimed. (You can do something similar if \( b \neq 0 \). Note there are many points on the plane. This just picks out one.)

The converse follows from the above discussion involving the point and a normal vector. This proves the proposition.

**Example 3.1.4** Find a normal vector to the plane \( 2x + 5y - z = 12.3 \).

A normal vector is \( (2, 5, -1) \). A point on this plane is \( (0, 0, -12.3) \). Of course there are many other points on this plane.

### 3.1.2 The Plane Which Contains Three Points

Sometimes you need to find the equation of a plane which contains three points. Consider the following picture.
3.1. FINDING PLANES

You have plenty of points but you need a normal. This can be obtained by taking \( \mathbf{a} \times \mathbf{b} \) where \( \mathbf{a} = (a_1 - a_0, b_1 - b_0, c_1 - c_0) \) and \( \mathbf{b} = (a_2 - a_0, b_2 - b_0, c_2 - c_0) \).

**Example 3.1.5** Find the equation of the plane which contains the three points, \((1, 2, 1), (3, -1, 2), \) and \((4, 2, 1)\).

You just need to get a normal vector to this plane. This can be done by taking the cross products of the two vectors, \((3, -1, 2) - (1, 2, 1)\) and \((4, 2, 1) - (1, 2, 1)\).

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 \(y + 3z = 5\). When you have what you think is the plane containing the three points, you ought to check it by seeing if it really does contain the three points.

**Example 3.1.6** 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

\[
\begin{vmatrix}
3 & 0 & 0 \\
2 & -3 & 1 \\
x - 1 & y - 2 & z - 1
\end{vmatrix} = 0.
\]

Hence \(-9z + 15 - 3y = 0\) and dividing by 3 yields the same answer as the above.

**Example 3.1.7** 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 proceed as above.

**Example 3.1.8** 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.
3.1.3 The Cartesian Equation For Lines

Recall the parametric equation of a line containing the point \( u \) and with direction vector \( v \) is

\[ u + tv, \quad t \in \mathbb{R} \]

For example, you can find a parametric equation of the line through the points \((1, 2, 3)\) and \((2, 1, 1)\). A direction vector for this line is \((1, -1, -2)\) and so a parametric equation of the line is

\[ (1, 2, 3) + t (1, -1, -2) = (1 + t, 2 - t, 3 - 2t) \]

Thus
\[ x = 1 + t, \quad y = 2 - t, \quad z = 3 - 2t \]

Solving each of these equations for \( t \) gives
\[ t = x - 1 = 2 - y = (3 - z) / 2 \]

Then the Cartesian equation of the line is
\[ x - 1 = 2 - y = (3 - z) / 2 \]

Note how there are two equations to determine the parametric equation of a line.

3.1.4 Parametric Equation For A Plane

Now consider a parametric equation for a plane. There exists such a thing and it is very similar to a parametric equation for a line. However, in the case of a plane, there are two parameters. I will illustrate with the following example.

**Example 3.1.9** Write a parametric equation for the plane whose Cartesian equation is

\[ x + 2y - z = 7 \]

What you do is to let \( y = s \) and \( z = t \). Then you have

\[ \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 7 - 2s + t \\ s \\ t \end{pmatrix} \]

You could also write it as the three equations
\[ x = 7 - 2s + t \]
\[ y = s \]
\[ z = t \]

In general, the parametric equation of a plane is of the form
\[ \overrightarrow{OP} = \overrightarrow{OP_0} + su + tv \]
3.1. FINDING PLANES

where $P$ is an arbitrary point on this plane and $\mathbf{u}, \mathbf{v}$ are two fixed vectors which are not parallel. Written in terms of components, this is of the form

$$
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} + s \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} + t \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}, \quad s, t \in \mathbb{R}
$$

For example

$$
\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + s \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + t \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \quad s, t \in \mathbb{R}
$$

is a parametric equation for a plane. To see geometrically why this does trace out a plane, consider first the case where $P_0 = 0$ so that the equation is of the form

$$
\begin{pmatrix} x \\ y \\ z \end{pmatrix} = s \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} + t \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = s\mathbf{u} + t\mathbf{v}, \quad s, t \in \mathbb{R}
$$

Here is a picture of the type of thing this describes.

When you add in $P_0 \neq 0$, it just takes all the points described by the above plane and adds the fixed vector $\overrightarrow{OP_0}$ to them. This gives another plane which goes through the point $P_0$ rather than the point $0$ but it will be geometrically parallel to the one just drawn.

Perhaps a better way to see the above gives the equation of a plane is to show it reduces to a Cartesian equation for a plane. It turns out that if $\mathbf{u}, \mathbf{v}$ are not parallel, then there will exist two of the equations in $\mathbb{R}^3$ which can be solved uniquely for $s$ and $t$ in terms two of the
Cartesian variables \( x, y, z \). Then this is substituted back in to the third equation and this gives a Cartesian equation for the plane. For example, consider \( 3.1 \). The last two equations give

\[
y = 2s + t, \quad z - 2 = 3s
\]

Therefore,

\[
s = \frac{z - 2}{3}
\]

And now from the first equation,

\[
y = 2 \left( \frac{z - 2}{3} \right) + t
\]

so solving for \( t \) yields

\[
t = y - \frac{2}{3}z + \frac{4}{3}
\]

Now substitute this in to the top equation.

\[
x = 1 + s + -t = 1 + \left( \frac{z - 2}{3} \right) - \left( y - \frac{2}{3}z + \frac{4}{3} \right)
\]

and if you simplify this you get

\[
x = -1 + z - y
\]

so a Cartesian equation for this plane is

\[
x + y = z = -1
\]

### 3.1.5 Triple Vector Product

Note that \((i \times j) \times j = -i\) and \(i \times (j \times j) = 0\) and so the cross product is not associative. Thus it makes absolutely \textbf{NO SENSE} to write \(a \times b \times c\) because the answer can depend on which cross product you do first. Thus

\[
a \times b \times c = \text{Garbage}
\]

However, it is often the case that you have to consider an expression of the form

\[
u \times (v \times w)
\]

This is called the triple vector product. Note the order in which you take the cross product is specified so there is no ambiguity. There is a nice reduction identity which simplifies this ugly expression. It is the following.

\[
u \times (v \times w) = (u \cdot w) v - (u \cdot v) w
\]

This identity follows directly from the coordinate description of the dot and cross products. For a systematic presentation of such vector identities, see the section on identities, Section \(2.3.2\). There is a convenient notation which allows you to discover all of these and you don’t have to memorize any of them.
3.1.6 The Angle Between Two Planes

Definition 3.1.10 Suppose two planes intersect. The angle between the planes is defined to be the angle between their normal vectors.

Example 3.1.11 Find the angle between the two planes, \( x + 2y - z = 6 \) and \( 3x + 2y - z = 7 \).

The two normal vectors are \((1, 2, -1)\) and \((3, 2, -1)\). Therefore, the cosine of the angle desired is

\[
\cos \theta = \frac{(1, 2, -1) \cdot (3, 2, -1)}{\sqrt{1^2 + 2^2 + (-1)^2} \sqrt{3^2 + 2^2 + (-1)^2}} = 0.87287
\]

Now use a calculator or table to find what the angle is. \( \cos \theta = 0.87287 \), Solution is \( \{ \theta = 0.50974 \} \). This value is in radians.

3.1.7 Intercepts Of A Plane

One way to understand how a plane looks is to connect the points where it intersects 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 3.1.12 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 3.1.13 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.

In general, to find the intercepts of a plane of the form \(ax + by + cz = d\) where \(d \neq 0\) and none of \(a, b,\) or \(c\) are equal to 0, divide by \(d\). This gives

\[
\frac{x}{(d/a)} + \frac{y}{(d/b)} + \frac{z}{(d/c)} = 1
\]

the intercepts are \((\frac{d}{a}, 0, 0)\), \((0, \frac{d}{b}, 0)\), \((0, 0, \frac{d}{c})\).
3.1.8 Distance Between A Point And A Plane Or A Point And A Line

There exists a stupid formula for the distance between a point and a plane. I will first illustrate with an example.

**Example 3.1.14** Find the distance from the point \((1, 2, 3)\) to the plane \(x - y + z = 3\).

The distance is the length of the line segment normal to the plane which goes from the given point to the given plane. In this example, a direction vector for this line is \((1, -1, 1)\), a normal vector to the plane. Thus the equation for the desired line is

\[
(x, y, z) = (1, 2, 3) + t(1, -1, 1)
\]

Lets find the value of \(t\) at which the line intersects the plane. Thus

\[
(1 + t) - (2 - t) + (3 + t) = 3
\]

and so \(t = \frac{1}{3}\). Therefore, the line segment is the one which joins \((1, 2, 3)\) to

\[
\begin{pmatrix}
\frac{4}{3} \\
\frac{5}{3} \\
\frac{10}{3}
\end{pmatrix}
\]

Now it follows the desired distance is

\[
\sqrt{\left(1 - \frac{4}{3}\right)^2 + \left(2 - \frac{5}{3}\right)^2 + \left(3 - \frac{10}{3}\right)^2} = \frac{1}{3}\sqrt{3}
\]

In the general case there is a simple and interesting geometrical consideration which will lead to a stupid formula which you can then use with no thought to do an uninteresting task, finding the distance from a point to a plane.

**Example 3.1.15** Find the distance from the point \((x_0, y_0, z_0)\) to the plane \(ax + by + cz = d\). Here \((a, b, c) \neq (0, 0, 0)\).

Consider the following picture in which \(P_0\) is a point in the plane and \(X_0 = (x_0, y_0, z_0)\) is the point whose distance to the plane is to be found. The normal to the plane is \(n\).

\[
\frac{|X_0 - P_0| \cos \theta}{|n|} = \frac{|X_0 - P_0| \cdot n}{|X_0 - P_0| \cdot |n|} = \frac{(X_0 - P_0) \cdot n}{|n|}
\]
3.1. FINDING PLANES

As drawn in the picture, \(|\mathbf{X}_0 - \mathbf{P}_0|\cos \theta\) will be positive but if you had \(\mathbf{n}\) pointing the opposite direction this would be negative. However, either way, it’s absolute value would give the right answer. This is why the absolute value is taken in the above. From this the stupid formula will follow easily. Suppose \(a \neq 0\). Things work the same if \(b\) or \(c\) are not zero. Then as explained above, you can take \(\mathbf{P}_0 = \left(\frac{d}{a}, 0, 0\right)\) and \(\mathbf{n} = (a, b, c)\). Therefore, the above expression is

\[
\left| \left( x_0 - \frac{d}{a}, y_0, z_0 \right) \cdot \frac{(a, b, c)}{\sqrt{a^2 + b^2 + c^2}} \right| = \frac{|ax_0 + by_0 + cz_0 - d|}{\sqrt{a^2 + b^2 + c^2}}
\]

and it is this last expression which is the stupid formula. Here is the same example done in an ad hoc manner earlier but this time through the use of a stupid formula.

Example 3.1.16 Find the distance from the point \((1, 2, 3)\) to the plane \(x - y + z = 3\).

Let’s apply the stupid formula. \(a = 1, b = -1, c = 1, d = 3, x_0 = 1, y_0 = 2, z_0 = 3\). Then plugging in to the formula, you get

\[
\frac{|1 \times 1 + (-1) \times 2 + 1 \times 3 - 3|}{\sqrt{1 + 1 + 1}} = \frac{1}{3}\sqrt{3}
\]

which gives the same answer much more easily. Those of you who expect to find the distance from a given point to a plane repeatedly, should certainly cherish and memorize this formula because it will save you lots of time. The rest of you should try to understand its derivation which is genuinely interesting and worth while. Unfortunately, finding the distance from a point to a plane is an excellent test question.

A similar formula holds for the distance from a point to a line in \(\mathbb{R}^2\). Recall from high school algebra, a line can be written as

\[ax + by = c\]

Then if \((x_0, y_0)\) is a point and you want the distance from this point to the given line, it equals

\[
\frac{|ax_0 + by_0 - c|}{\sqrt{a^2 + b^2}}.
\]

You should derive this stupid formula from the same geometric considerations used to get the stupid formula for a point and a plane.

Of course it all generalizes. The same reasoning will yield a stupid formula for the distance between \((y_1, \cdots, y_n)\) and the level surface, called a hyperplane given by \(\sum_{k=1}^{n} a_k x_k = d\). You can probably guess what it is by analogy to the above but it is better to derive it directly using the same sort of geometric reasoning just given.

\[\text{Words such as “hyper” give an aura of significance to things which are in reality trivial while obfuscating the real issues. They constitute an example of pretentious jargon which militates against correct understanding.}\]
Chapter 4

Systems Of Linear Equations

Quiz

1. The intercepts of a plane are \((1, 0, 0), (0, -1, 0), \) and \((0, 0, 1)\). Find the equation of the plane.

2. A plane has a normal vector \((1, 0, -3)\) and contains the point \((1, 1, 2)\). Find the equation of the plane.

3. Find the equation of a plane which has the three points, \((1, 1, 1), (2, -2, 1), (0, 3, 0)\).

4. Find the cross product, \(i - j + k \times i + k\). What is the area of the triangle determined by these two vectors.

4.1 Systems Of Equations, Geometric Interpretations

As you know, 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.

![Graph showing one solution, two parallel lines, no solutions, and infinitely many solutions.]

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 4.1.1 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 4.1.2 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.

The points, \((x, y, z)\) satisfying an equation in three variables like \(2x + 4y - 5z = 8\) form a plane in three dimensions and geometrically, when you solve systems of equations involving three variables, you are taking intersections of planes. Consider the following picture involving two planes.
Notice how these two planes intersect in a line. It could also happen the two planes could fail to intersect.

Now imagine a third plane. One thing that could happen is this third plane could have an intersection with one of the first planes which results in a line which fails to intersect the first line as illustrated in the following picture.

Thus there is no point which lies in all three planes. The picture illustrates the situation in which the line of intersection of the new plane with one of the original planes forms a line parallel to the line of intersection of the first two planes. However, in three dimensions, it is possible for two lines to fail to intersect even though they are not parallel. Such lines are called skew lines. You might consider whether there exist two skew lines, each of which is the intersection of a pair of planes selected from a set of exactly three planes such that there is no point of intersection between the three planes. You can also see that if you tilt one of the planes you could obtain every pair of planes having a nonempty intersection in a line and yet there may be no point in the intersection of all three.

It could happen also that the three planes could intersect in a single point as shown in the following picture.

In this case, the three planes have a single point of intersection. The three planes could
also intersect in a line.

Thus in the case of three equations having three variables, the planes determined by these equations could intersect in a single point, a line, or even fail to intersect at all. You see that in three dimensions there are many possibilities. If you want to waste some time, you can try to imagine all the things which could happen but this will not help for dimensions higher than 3 which is where many of the important applications lie.

In higher dimensions it is customary to refer to the set of points described by relations like $x + y - 2z + 4w = 8$ as hyper-planes. 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. The only rational and useful way to deal with this subject is through the use of algebra. Indeed, a major reason for studying mathematics is to obtain freedom from always having to draw a picture in order to do a computation or find out something important.

4.2 Systems Of Equations, Algebraic Procedures

4.2.1 Elementary Operations

Definition 4.2.1 A system of linear equations is a set of $p$ equations for the $n$ variables, $x_1, \ldots, x_n$ which is of the form

$$\sum_{k=1}^{n} a_{mk}x_k = d_m, m = 1, 2, \ldots, p$$

Written less compactly it is a set of equations of the following form

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = d_1$$
$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = d_2$$
$$\vdots$$
$$a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pn}x_n = d_p$$

The evocative semi word, “hyper” conveys absolutely no meaning but is traditional usage which makes the terminology sound more impressive than something like long wide comparatively flat thing which does convey some meaning. However, in such cases as these pretentious jargon is nearly always preferred. Later we will discuss some terms which are not just evocative but yield real understanding.
The problem is to find the values of $x_1, x_2, \cdots, x_n$ which satisfy all $p$ equations. This is called the solution set of the system of equations. In other words, $(a_1, \cdots, a_n)$ is in the solution set of the system of equations if when you plug $a_1$ in place of $x_1$, $a_2$ in place of $x_2$ etc., each equation in the system is satisfied.

Consider the following example.

**Example 4.2.2** Find $x$ and $y$ such that

$$x + y = 7 \quad \text{and} \quad 2x - y = 8.$$  \hfill (4.1)

The set of ordered pairs, $(x, y)$ which solve both equations is called the 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 4.2.3** 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 4.2.4** To illustrate the third of these operations on this particular system, consider the following.

$$x + y = 7$$
$$2x - y = 8$$

The system has the same solution set as the system

$$x + y = 7$$
$$-3y = -6.$$  \hfill (4.2)

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

$$-3y = -6.$$  \hfill (4.3)

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 4.2.3 do not change the set of solutions to the system of linear equations.

**Theorem 4.2.5** Suppose you have two equations, involving the variables, $(x_1, \cdots, x_n)$

$$E_1 = f_1, \quad E_2 = f_2$$  \hfill (4.2)

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

$$E_1 = f_1, \quad E_2 + aE_1 = f_2 + af_1.$$  \hfill (4.3)

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 \(E_1 = f_1, E_2 = f_2 + af_1\). 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 \(E_1 = f_1, E_2 = af_2\) provided \(a \neq 0\) is verified as follows: If \((x_1,\ldots,x_n)\) 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 \((x_1,\ldots,x_n)\) 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 4.2.6** Find the solutions to the system,

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

(4.4)

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

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

(4.5)

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

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

(4.6)

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

Alternatively, in (4.6) 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

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

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

### 4.2.2 Gauss Elimination, Row Echelon Form

A less cumbersome way to represent a linear system is to write it as an augmented matrix. For example the linear system, \(4.4\) 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 \(4.4\) 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 \(4.5\). 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 \(4.6\). By back substitution you obtain the solution \(x = 1, y = 6,\) and \(z = 3.\)
In general a linear system is of the form

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

(4.7)

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

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

(4.8)

Changes to the system of equations in (4.7) as a result of an elementary operations translate into changes of the augmented matrix resulting from a row operation. Note that Theorem 4.2.5 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 4.2.7** 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 4.2.8** An augmented matrix is in **echelon form** also called **row 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.

**Definition 4.2.9** An augmented matrix is in **row reduced echelon form** if

1. All nonzero rows are above any rows of zeros.
2. Each leading entry of a row is in a column to the right of the leading entries of any rows above it.
3. All entries in a column above and below a leading entry are zero.
4. Each leading entry is a 1, the only nonzero entry in its column.

The relation between these two definitions is as described in the following picture.
Thus if the matrix is in row reduced echelon form, it is in row echelon form but not necessarily the other way around. You can usually find the solution to a system of equations by row reducing to row echelon form. You typically don’t have to go all the way to the row reduced echelon form but the row reduced echelon form is very important because, unlike a row echelon form, it is unique. It is also easier to use in the case where the system of equations has an infinite solution set.

**Example 4.2.10** Here are some augmented matrices which are in row reduced echelon form.

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

**Example 4.2.11** Here are augmented matrices in echelon form which are not in row reduced echelon form but which are in echelon form.

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

**Example 4.2.12** Here are some augmented matrices which are not in echelon form.

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

**Definition 4.2.13** 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 4.2.14** 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 following is the algorithm for obtaining a matrix which is in row reduced echelon form.

**Algorithm 4.2.15**

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

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

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

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

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

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

This row reduction procedure applies to both augmented matrices and non augmented matrices. There is nothing special about the augmented column with respect to the row reduction procedure.
Example 4.2.16  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

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

Now the algorithm will be applied to the matrix,

\[
\begin{bmatrix}
0 \\
-3 \\
\end{bmatrix}
\]

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.

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

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.

\[
\begin{bmatrix}
0 & 1 & 1 & 4 & 0 \\
0 & 0 & 6 & 9 & 0 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]

Next work on the fourth column in the same way.

\[
\begin{bmatrix}
0 & 3 & 3 & 0 & 0 \\
0 & 0 & 6 & 0 & 0 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]
Take \(-\frac{1}{2}\) times the second row and add to the first.

\[
\begin{pmatrix}
0 & 3 & 0 & 0 & 0 \\
0 & 0 & 6 & 0 & 0 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}.
\]

Finally, divide by the value of the leading entries in the nonzero rows.

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

The above algorithm is the way a computer would obtain a reduced echelon form for a given matrix. It is not necessary for you to pretend you are a computer but if you like to do so, the algorithm described above will work. The main idea is to do row operations in such a way as to end up with a matrix in echelon form or row reduced echelon form because when this has been done, the resulting augmented matrix will allow you to describe the solutions to the linear system of equations in a meaningful way.

**Example 4.2.17** 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 4.2.18** 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
\[
\begin{pmatrix}
3 & -1 & -5 & | & 9 \\
0 & 1 & -10 & | & 0 \\
0 & 0 & 0 & | & 0
\end{pmatrix}
\]

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

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

Take the middle row and add to the top and then divide the top row which results by 3.

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

This is in reduced echelon form. The equations corresponding to this reduced echelon form are \( y = 10z \) 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 **basic variables** or **leading variables**. The other variables are called **free variables**. This is because in solving the system, these variables are “free” to be anything. In Example 4.2.18 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 4.2.18 this parameter was \( t \). Sometimes there are many free variables and in these cases, you need to use many parameters. Here is another example.

\(^2\)In this context \( t \) is called a **parameter**.
Example 4.2.19 Find the solution to the system

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

The augmented matrix is

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

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

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

Now add the second row to the bottom row

\[
\begin{pmatrix}
  1 & 2 & -1 & 1 & | & 3 \\
  0 & -1 & 0 & 0 & | & -2 \\
  0 & 0 & 0 & 0 & | & 0
\end{pmatrix}
\]  \hfill (4.9)

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

\[
\begin{pmatrix}
  x \\
  y \\
  z \\
  w
\end{pmatrix} = \begin{pmatrix}
  -1 + s - t \\
  2 \\
  s \\
  t
\end{pmatrix}.
\]  \hfill (4.10)

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 row reduced echelon form rather than just echelon form before seeking to write down the description of the solution. In the above, this means we don’t stop with the echelon form 4.9. 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 4.10.

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
\[
\begin{pmatrix}
1 & 2 & \vdots & 3 \\
0 & 4 & \vdots & -2 \\
0 & 0 & \vdots & 1 \\
\end{pmatrix}
\]
or there is a unique solution as in the case where row operations yield an echelon form like
\[
\begin{pmatrix}
1 & 2 & 2 & \vdots & 3 \\
0 & 4 & 3 & \vdots & -2 \\
0 & 0 & 4 & \vdots & 1 \\
\end{pmatrix}.
\]
Also, sometimes there are free variables and no solution as in the following:
\[
\begin{pmatrix}
1 & 2 & 2 & \vdots & 3 \\
0 & 4 & 3 & \vdots & -2 \\
0 & 0 & 0 & \vdots & 1 \\
\end{pmatrix}.
\]
There are a lot of cases to consider but it is not necessary to make a major production of this. Do row operations till you obtain a matrix in echelon form or reduced echelon form and determine whether there is a solution. If there is, see if there are free variables. In this case, there will be infinitely many solutions. Find them by assigning different parameters to the free variables and obtain the solution. If there are no free variables, then there will be a unique solution which is easily determined once the augmented matrix is in echelon or row reduced echelon form. In every case, the process yields a straightforward way to describe the solutions to the linear system. As indicated above, you are probably less likely to become confused if you place the augmented matrix in row reduced echelon form rather than just echelon form.

In summary,

**Definition 4.2.20** A system of linear equations is a list of equations,
\[
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\
\vdots \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m
\]
where \(a_{ij}\) are numbers, and \(b_j\) is a number. The above is a system of \(m\) equations in the \(n\) variables, \(x_1, x_2, \ldots, x_n\). Nothing is said about the relative size of \(m\) and \(n\). Written more simply in terms of summation notation, the above can be written in the form
\[
\sum_{j=1}^{n} a_{ij}x_j = f_j, \quad i = 1, 2, 3, \ldots, m
\]
It is desired to find \((x_1, \ldots, x_n)\) 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.
Definition 4.2.21 A system of linear equations is called consistent if there exists a solution. It is called inconsistent if there is no solution.

These are reasonable words to describe the situations of having or not having a solution. If you think of each equation as a condition which must be satisfied by the variables, consistent would mean there is some choice of variables which can satisfy all the conditions. Inconsistent means there is no choice of the variables which can satisfy each of the conditions.
Chapter 5

Matrices

5.1 Matrix Operations And Algebra

Quiz

1. Here are three points: (1,1,1), (2,0,1), (0,1,0). Find an equation of a plane which contains all three points.

2. Find the equation of a plane which is parallel to the plane whose equations is \( x + 2y + z = 7 \) which contains the point (1, 2, 1).

3. Here are three vectors: (1,2,1), (2,1,0), (-2,0,1). Find the volume of the parallelepiped determined by these three vectors.

4. Here is a system of equations.

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

Find the complete solution.

5.1.1 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

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

because they are different sizes. As noted above, you write \((c_{ij})\) for the matrix \(C\) whose \(ij^{th}\) entry is \(c_{ij}\). In doing arithmetic with matrices you must define what happens in terms of the \(c_{ij}\) sometimes called the **entries** of the matrix or the **components** of the matrix.

The above discussion stated for general matrices is given in the following definition.

**Definition 5.1.1** (Scalar Multiplication) If \(A = (a_{ij})\) and \(k\) is a scalar, then \(kA = (ka_{ij})\).

**Example 5.1.2** \(7 \begin{pmatrix} 2 & 0 \\ 1 & -4 \end{pmatrix} = \begin{pmatrix} 14 & 0 \\ 7 & -28 \end{pmatrix}\).

**Definition 5.1.3** (Addition) If \(A = (a_{ij})\) and \(B = (b_{ij})\) are two \(m \times n\) matrices. Then \(A + B = C\) where

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

for \(c_{ij} = a_{ij} + b_{ij}\).
5.1. MATRIX OPERATIONS AND ALGEBRA

Example 5.1.4
\[
\begin{pmatrix}
1 & 2 & 3 \\
1 & 0 & 4
\end{pmatrix} + \begin{pmatrix}
5 & 2 & 3 \\
-6 & 2 & 1
\end{pmatrix} = \begin{pmatrix}
6 & 4 & 6 \\
-5 & 2 & 5
\end{pmatrix}
\]

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

Definition 5.1.5 (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 5.1.6 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 5.1.7 (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.
  \[ A + B = B + A, \] (5.1)
- Associative Law for Addition.
  \[ (A + B) + C = A + (B + C), \] (5.2)
- Existence of an Additive Identity
  \[ A + 0 = A, \] (5.3)
- Existence of an Additive Inverse
  \[ A + (-A) = 0, \] (5.4)

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

- Distributive law over Matrix Addition.
  \[ \alpha (A + B) = \alpha A + \alpha B, \] (5.5)
- Distributive law over Scalar Addition
  \[ (\alpha + \beta) A = \alpha A + \beta A, \] (5.6)
- Associative law for Scalar Multiplication
  \[ \alpha (\beta A) = \alpha \beta (A), \] (5.7)
- Rule for Multiplication by 1
  \[ 1A = A. \] (5.8)

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.
5.1.2 Multiplication Of Matrices

Definition 5.1.8 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,

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

Another way to think of this is

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

Thus you take $x_1$ times the first column, add to $x_2$ times the second column, and finally $x_3$ times the third column. In general, here is the definition of how to multiply an $(m \times n)$ matrix times a $(n \times 1)$ matrix.

Definition 5.1.9 Let $A = A_{ij}$ be an $m \times n$ matrix and let $\mathbf{v}$ be an $n \times 1$ matrix,

$$\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$
5.1. MATRIX OPERATIONS AND ALGEBRA

Then \( Av \) is an \( m \times 1 \) matrix and the \( i^{th} \) component of this matrix is

\[
(Av)_i = A_{i1}v_1 + A_{i2}v_2 + \cdots + A_{in}v_n = \sum_{j=1}^{n} A_{ij}v_j.
\]

Thus

\[
Av = \begin{pmatrix}
\sum_{j=1}^{n} A_{1j}v_j \\
\vdots \\
\sum_{j=1}^{n} A_{mj}v_j
\end{pmatrix}.
\] (5.9)

In other words, if

\[
A = (a_1, \ldots, a_n)
\]

where the \( a_k \) are the columns,

\[
Av = \sum_{k=1}^{n} v_k a_k
\]

This follows from 5.9 and the observation that the \( j^{th} \) column of \( A \) is

\[
\begin{pmatrix}
A_{1j} \\
A_{2j} \\
\vdots \\
A_{mj}
\end{pmatrix}
\]

so 5.9 reduces to

\[
v_1 \begin{pmatrix}
A_{11} \\
A_{21} \\
\vdots \\
A_{m1}
\end{pmatrix} + v_2 \begin{pmatrix}
A_{12} \\
A_{22} \\
\vdots \\
A_{m2}
\end{pmatrix} + \cdots + v_n \begin{pmatrix}
A_{1n} \\
A_{2n} \\
\vdots \\
A_{mn}
\end{pmatrix}
\]

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 5.1.10 Compute

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

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
\[
\begin{pmatrix}
1 & 2 & 1 & 3 \\
0 & 2 & 1 & -2 \\
2 & 1 & 4 & 1
\end{pmatrix}
\begin{pmatrix}
1 \\
2 \\
0 \\
1
\end{pmatrix}
= 
\begin{pmatrix}
8 \\
2 \\
5
\end{pmatrix}.
\]

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 5.1.11** 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

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

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

\[
AB \equiv (Ab_1, \ldots, Ab_p)
\]

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

**Example 5.1.12** 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

\[
\begin{pmatrix}
\text{First column} & \text{Second column} & \text{Third column}
\end{pmatrix}
\begin{pmatrix}
1 & 1 \\
2 & 2 \\
1 & 0
\end{pmatrix}
\begin{pmatrix}
2 \\
3 \\
1
\end{pmatrix}
\begin{pmatrix}
0 \\
1 \\
1
\end{pmatrix}
\]

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 5.1.13 Multiply the following.

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

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!

5.1.3 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

\[
\sum_{k=1}^{n} A_{ik}B_{kj}. \tag{5.11}
\]

In terms of pictures of the matrix, you are doing

\[
\begin{pmatrix}
A_{11} & A_{12} & \cdots & A_{1n} \\
A_{21} & A_{22} & \cdots & A_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
A_{m1} & A_{m2} & \cdots & A_{mn}
\end{pmatrix}
\begin{pmatrix}
B_{11} & B_{12} & \cdots & B_{1p} \\
B_{21} & B_{22} & \cdots & B_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
B_{n1} & B_{n2} & \cdots & B_{np}
\end{pmatrix}
\]

Then as explained above, the \(j\)th column is of the form

\[
\begin{pmatrix}
A_{11} & A_{12} & \cdots & A_{1n} \\
A_{21} & A_{22} & \cdots & A_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
A_{m1} & A_{m2} & \cdots & A_{mn}
\end{pmatrix}
\begin{pmatrix} B_{1j} \\ B_{2j} \\ \vdots \\ B_{nj} \end{pmatrix}
\]

which is a \(m \times 1\) matrix or column vector which equals

\[
\begin{pmatrix}
A_{11} \\
A_{21} \\
\vdots \\
A_{m1}
\end{pmatrix} B_{1j} + \begin{pmatrix}
A_{12} \\
A_{22} \\
\vdots \\
A_{m2}
\end{pmatrix} B_{2j} + \cdots + \begin{pmatrix}
A_{1n} \\
A_{2n} \\
\vdots \\
A_{mn}
\end{pmatrix} B_{nj}.
\]
The second entry of this $m \times 1$ matrix is

$$A_{21}B_{1j} + A_{22}B_{2j} + \cdots + A_{2n}B_{nj} = \sum_{k=1}^{n} A_{2k}B_{kj}.$$ 

Similarly, the $i^{th}$ entry of this $m \times 1$ matrix is

$$A_{i1}B_{1j} + A_{i2}B_{2j} + \cdots + A_{in}B_{nj} = \sum_{k=1}^{n} A_{ik}B_{kj}.$$ 

This shows the following definition for matrix multiplication in terms of the $ij^{th}$ entries of the product coincides with Definition 5.1.11.

**Definition 5.1.14** 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

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}. \quad (5.12)$$

*Note this says that to find $(AB)_{ij}$, you multiply the $i^{th}$ row of $A$ on the left by the $j^{th}$ column of $B$ on the right.*

**(Example 5.1.15)** 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

$$\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 2 \\ 7 \\ 6 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 3 \\ 6 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

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 $3 \times 3$ 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

$$\sum_{j} a_{3k}b_{kj} = a_{31}b_{12} + a_{32}b_{22} = 2 \times 3 + 6 \times 6 = 42.$$ 

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 5.1.16 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 5.1.17 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 5.1.18 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\[\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 2 & 4 & 2 & 0 \\ 1 & 2 & 1 & 0 \end{pmatrix} \].

5.1.4 Graphs And Digraphs

Consider the following picture. It is referred to as a digraph or directed graph.

There are three locations in this graph, labelled 1, 2, and 3. The directed lines represent a way of going from one location to another. Thus there is one way to go from location 1 to location 1. There is one way to go from location 1 to location 3. It is not possible to go
from location 2 to location 3 although it is possible to go from location 3 to location 2. Let’s refer to moving along one of these directed lines as a step. The following $3 \times 3$ matrix is a numerical way of writing the information in the above picture. When you have one of these things in which there you can go from $i$ to $j$ if and only if you can go from $j$ to $i$, then it is called a graph. Thus the matrix of a graph must be symmetric. This is not the case for the above picture.

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

Thus $a_{ij}$, the entry in the $i^{th}$ row and $j^{th}$ column represents the number of ways to go from location $i$ to location $j$ in one step.

**Problem:** Find the number of ways to go from $i$ to $j$ using exactly $k$ steps.

Denote the answer to the above problem by $a_{ij}^k$. We don’t know what it is right now unless $k = 1$ when it equals $a_{ij}$ described above. However, if we did know what it was, we could find $a_{ij}^{k+1}$ as follows.

$$
a_{ij}^{k+1} = \sum_r a_{ir}^k a_{rj}
$$

This is because if you go from $i$ to $j$ in $k + 1$ steps, you first go from $i$ to $r$ in $k$ steps and then for each of these ways there are $a_{rj}$ ways to go from there to $j$. Thus $a_{ir}^k a_{rj}$ gives the number of ways to go from $i$ to $j$ in $k + 1$ steps such that the $k^{th}$ step leaves you at location $r$. Adding these gives the above sum. Now you recognize this as the $ij^{th}$ entry of the product of two matrices. Thus

$$
a_{ij}^2 = \sum_r a_{ir} a_{rj}
$$

and so forth. From the above definition of matrix multiplication, Definition 5.1.14, this shows that if $A$ is the matrix associated with the directed graph as above, then $a_{ij}^k$ is just the $ij^{th}$ entry of $A^k$ where $A^k$ is just what you would think it should be, $A$ multiplied by itself $k$ times.

Thus in the above example, to find the number of ways of going from 1 to 3 in two steps you would take that matrix and multiply it by itself and then take the entry in the first row and third column. Thus

$$
\begin{pmatrix}
1 & 1 & 1 \\
1 & 0 & 0 \\
1 & 1 & 0
\end{pmatrix}^2 = \begin{pmatrix}
3 & 2 & 1 \\
1 & 1 & 1 \\
2 & 1 & 1
\end{pmatrix}
$$

and you see there is exactly one way to go from 1 to 3 in two steps. You can easily see this is true from looking at the graph also. Note there are three ways to go from 1 to 1 in 2 steps. Can you find them from the picture? What would you do if you wanted to consider 5 steps?

$$
\begin{pmatrix}
1 & 1 & 1 \\
1 & 0 & 0 \\
1 & 1 & 0
\end{pmatrix}^5 = \begin{pmatrix}
28 & 19 & 13 \\
13 & 9 & 6 \\
19 & 13 & 9
\end{pmatrix}
$$

There are 19 ways to go from 1 to 2 in five steps. Do you think you could list them all by looking at the picture? I don’t think you could do it without wasting a lot of time.
5.1. MATRIX OPERATIONS AND ALGEBRA

Of course there is nothing sacred about having only three locations. Everything works just as well with any number of locations. In general if you have \( n \) locations, you would need to use a \( n \times n \) matrix.

**Example 5.1.19** Consider the following directed graph.

Write the matrix which is associated with this directed graph and find the number of ways to go from 2 to 4 in three steps.

Here you need to use a \( 4 \times 4 \) matrix. The one you need is

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

Then to find the answer, you just need to multiply this matrix by itself three times and look at the entry in the second row and fourth column.

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

There is exactly one way to go from 2 to 4 in three steps.

How many ways would there be of going from 2 to 4 in five steps?

\[
\begin{pmatrix}
0 & 1 & 1 & 0 \\
1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 \\
0 & 1 & 0 & 1
\end{pmatrix}^5 = \begin{pmatrix}
5 & 9 & 5 & 4 \\
5 & 4 & 1 & 3 \\
9 & 10 & 4 & 6 \\
4 & 6 & 3 & 3
\end{pmatrix}
\]

There are three ways. Note there are 10 ways to go from 3 to 2 in five steps. Would you care to find each of these by looking at the picture of the graph?

This is an interesting application of the concept of the \( ij^{th} \) entry of the product matrices.
5.1.5 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 5.1.20** Compare \[
\begin{pmatrix}
1 & 2 \\
3 & 4
\end{pmatrix}
\begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}
\]
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 \[
\begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}
\begin{pmatrix}
1 & 2 \\
3 & 4
\end{pmatrix}
= \begin{pmatrix}
3 & 4 \\
1 & 2
\end{pmatrix}
\].

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 5.1.21** If all multiplications and additions make sense, the following hold for matrices, \(A, B, C\) and \(a, b\) scalars.

\[
A (aB + bC) = a (AB) + b (AC)
\] (5.13)

\[
(B + C) A = BA + CA
\] (5.14)

\[
A (BC) = (AB) C
\] (5.15)

**Proof:** Using Definition 5.1.14,

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

Thus \(AB + AC\) as claimed. Formula 5.14 is entirely similar.

Formula 5.15 is the associative law of multiplication. Using Definition 5.1.4,

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

This proves 5.15.
5.1. MATRIX OPERATIONS AND ALGEBRA

5.1.6 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 5.1.22** 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 5.1.23**

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

The transpose of a matrix has the following important properties.

**Lemma 5.1.24** Let \( A \) be an \( m \times n \) matrix and let \( B \) be a \( n \times p \) matrix. Then

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

(5.16)

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

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

(5.17)

**Proof:** From the definition,

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

The proof of Formula (5.17) is left as an exercise and this proves the lemma.

**Definition 5.1.25** 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 5.1.26** Let

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

Then \( A \) is symmetric.
Example 5.1.27  Let

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

Then \( A \) is skew symmetric.

5.1.7  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.

\[
(1), \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}
\]

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 \( ij^{th} \) entry of the identity matrix

\[
I_{ij} = \delta_{ij}
\]

where \( \delta_{ij} \) is the Kroneker 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 5.1.28  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 5.1.29  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 5.1.30  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 5.1.31** 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
\[
\begin{pmatrix} 0 \\ 0 \end{pmatrix} = A^{-1} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = A^{-1} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix},
\]
a contradiction. Thus the answer is that \( A \) does not have an inverse.

**Example 5.1.32** 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 \( A \).

Here is a little proposition. First a definition is needed.

**Definition 5.1.33** Let \( A \) be an \( m \times n \) matrix. Thus multiplication on the left of a vector in \( \mathbb{F}^n \) yields a vector in \( \mathbb{F}^m \). Then \( A \) is said to be onto if for every \( y \in \mathbb{F}^m \) there exists \( x \in \mathbb{F}^n \) such that \( Ax = y \). The matrix is said to be one to one if whenever \( Ax = 0 \) it follows \( x = 0 \).

**Proposition 5.1.34** Let \( A \) be an \( n \times n \) matrix. Thus \( Ax = y \) where \( y, x \in \mathbb{F}^n \). Then \( A \) is invertible if and only if \( A \) is one to one and onto.

**Proof**: Suppose \( A \) is invertible. Why is \( A \) onto? Let \( y \in \mathbb{F}^n \) and consider \( A^{-1}y \). Then
\[
A \left( A^{-1}(y) \right) = (AA^{-1})y = Iy = y
\]
Thus $A$ is onto. Why is $A$ one to one? Suppose $Ax = 0$. Then multiplying on both sides by $A^{-1}$ yields
\[ Ix = A^{-1}Ax = A^{-1}0 = 0 \]
Now suppose $A$ is one to one and onto. Why is it invertible? Since $A$ is onto and one to one, there exists a unique $b_i$ such that
\[ Ab_i = e_i \]
where $e_i$ is the column vector which has a 1 in the $i^{th}$ slot and zeros everywhere else. Thus from Definition 5.1.11
\[ A \left( \begin{array}{c} b_1 \\ b_2 \\ \vdots \\ b_n \end{array} \right) = \left( \begin{array}{c} e_1 \\ e_2 \\ \vdots \\ e_n \end{array} \right) = I \]
Thus the matrix $B$ which has $b_i$ in the $i^{th}$ position is a candidate for the inverse. It works well on the right of $A$. Now define a function $L : \mathbb{F}^n \to \mathbb{F}^n$ by
\[ Ly \equiv x \]
where $Ax = y$. In other words $L$ sends $y$ to where it come from by way of $A$. Then this function is linear which means
\[ L(ay_1 + by_2) = aL(y_1) + bL(y_2) \]
Now
\[ c_j \equiv Le_j, \quad C \equiv \left( \begin{array}{c} c_1 \\ \vdots \\ c_n \end{array} \right) \]
Then
\[ L = C \]
because if you do both sides to $e_j$ you get $c_j$ from the definition of $C$. Since the two sides agree on the special vectors $e_k$, they must agree on all vectors because any vector $x$ can be written as
\[ x = \sum_{j=1}^{n} x_j e_j \]
and both sides are linear. Hence letting $id$ denote the identity map which sends every vector to itself,
\[ id = LA = CA \]
and so it follows that for all $x$,
\[ x = CAx \]
Hence $I = CA$. In particular
\[ e_k = CAe_k \]
which equals the $k^{th}$ column of $CA$. Thus $CA = I$ the identity matrix. Recall
\[ I = AB \]
Now multiply both sides on the left by $C$ and obtain
\[ C = C(AB) = (CA)B = IB = B \]
Hence $BA = AB = I$ and so $A$ is invertible.
5.2 Finding The Inverse Of A Matrix, Gauss Jordan Method

Quiz

1. Multiply the matrices if possible.

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

2. Multiply the matrices if possible.

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

3. Multiply the matrices if possible.

\[
\begin{pmatrix}
1 \\
0 \\
1 \\
1
\end{pmatrix}
\begin{pmatrix}
x & z \\
y & w
\end{pmatrix}
\]

4. True or False. In each case the capital letters are matrices of an appropriate size and the lower case letters represent numbers.

(a) \( A^2 - B^2 = (A - B)(A + B) \)
(b) \( (AB)^T = A^T B^T \)
(c) \( (aA + bB)C = aAC + bCB \)
(d) If \( AB = 0 \), then either \( A = 0 \) or \( B = 0 \).
(e) \( A/A = 1 \)
(f) \( (AB)C = A(BC) \)

In the Example 5.1.32, 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
\[
\begin{pmatrix}
1 & 1 & | & 1 \\
1 & 2 & | & 0
\end{pmatrix}
\] (5.18)
for the first system and
\[
\begin{pmatrix}
1 & 1 & | & 0 \\
1 & 2 & | & 1
\end{pmatrix}
\] (5.19)
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, 5.19 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
\[
\begin{pmatrix}
2 & -1 \\
-1 & 1
\end{pmatrix}
\].

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, \(\begin{pmatrix} x \\ y \end{pmatrix}\), the first
column of the inverse was obtained first and then the second column \(\begin{pmatrix} z \\ w \end{pmatrix}\).

To simplify this procedure, you could have written
\[
\begin{pmatrix}
1 & 1 & | & 1 & 0 \\
1 & 2 & | & 0 & 1
\end{pmatrix}
\]
and row reduced till you obtained
\[
\begin{pmatrix}
1 & 0 & | & 2 & -1 \\
0 & 1 & | & -1 & 1
\end{pmatrix}
\]
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 \textbf{Gauss-Jordan procedure}.
5.2. FINDING THE INVERSE OF A MATRIX, GAUSS JORDAN METHOD

Procedure 5.2.1 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

$$ (I|B). $$

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 5.2.2 Let $A = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 3 & 1 & -1 \end{pmatrix}$. Find $A^{-1}$ if it exists.

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

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

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

$$ \begin{pmatrix} 1 & 2 & 2 & | & 1 & 0 & 0 \\ 0 & -2 & 0 & | & -1 & 1 & 0 \\ 0 & -5 & -7 & | & -3 & 0 & 1 \end{pmatrix} $$

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

$$ \begin{pmatrix} 1 & 2 & 2 & | & 1 & 0 & 0 \\ 0 & -10 & 0 & | & -5 & 5 & 0 \\ 0 & 0 & 14 & | & 1 & 5 & -2 \end{pmatrix} $$

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

$$ \begin{pmatrix} 1 & 0 & 0 & | & -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\ 0 & 1 & 0 & | & \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & | & \frac{1}{14} & \frac{5}{14} & -\frac{1}{7} \end{pmatrix} $$
Therefore, the inverse is
\[
\left(\begin{array}{ccc}
-\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\
\frac{1}{2} & -\frac{1}{2} & 0 \\
\frac{1}{14} & \frac{5}{14} & -\frac{1}{7}
\end{array}\right)
\]

**Example 5.2.3** 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 5.2.4** 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.

**Example 5.2.5** In this example, it is shown how to use the inverse of a matrix to find the solution to a system of equations. Consider the following system of equations. Use the inverse of a suitable matrix to give the solutions to this system.
\[
\begin{pmatrix}
x + z = 1 \\
x - y + z = 3 \\
x + y - z = 2
\end{pmatrix}.
\]
The system of equations can be written in terms of matrices as
\[
\begin{pmatrix}
1 & 0 & 1 \\
1 & -1 & 1 \\
1 & 1 & -1
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= \begin{pmatrix}
1 \\
3 \\
2
\end{pmatrix}.
\] (5.21)
More simply, this is of the form $Ax = b$. Suppose you find the inverse of the matrix, $A^{-1}$. Then you could multiply both sides of this equation by $A^{-1}$ to obtain
\[
x = (A^{-1}A)x = A^{-1}(Ax) = A^{-1}b.
\]
This gives the solution as $x = A^{-1}b$. Note that once you have found the inverse, you can easily get the solution for different right hand sides without any effort. It is always just $A^{-1}b$. In the given example, the inverse of the matrix is
\[
\begin{pmatrix}
0 & \frac{1}{2} & \frac{1}{2} \\
1 & -1 & 0 \\
1 & -\frac{1}{2} & -\frac{1}{2}
\end{pmatrix}
\]
This was shown in Example 5.2.4. Therefore, from what was just explained the solution to the given system is
\[
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= \begin{pmatrix}
0 & \frac{1}{2} & \frac{1}{2} \\
1 & -1 & 0 \\
1 & -\frac{1}{2} & -\frac{1}{2}
\end{pmatrix}
\begin{pmatrix}
1 \\
3
\end{pmatrix}
= \begin{pmatrix}
\frac{5}{2} \\
-2 \\
-\frac{3}{2}
\end{pmatrix}.
\]
What if the right side of 5.21 had been
\[
\begin{pmatrix}
0 \\
1 \\
3
\end{pmatrix}?
What would be the solution to
\[
\begin{pmatrix}
1 & 0 & 1 \\
1 & -1 & 1 \\
1 & 1 & -1
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= 
\begin{pmatrix}
0 \\
1 \\
3
\end{pmatrix}?
\]

By the above discussion, it is just
\[
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= 
\begin{pmatrix}
0 & \frac{1}{2} & \frac{1}{2} \\
1 & -1 & 0 \\
1 & -\frac{1}{2} & -\frac{1}{2}
\end{pmatrix}
\begin{pmatrix}
0 \\
1 \\
3
\end{pmatrix} = 
\begin{pmatrix}
2 \\
-1 \\
-2
\end{pmatrix}.
\]

This illustrates why once you have found the inverse of a given matrix, you can use it to solve many different systems easily.

Here is a formula for the inverse of a 2×2 matrix.

**Theorem 5.2.6** Let \( A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \) where \( ad - bc \neq 0 \). Then

\[
A^{-1} = \frac{1}{ad - cb} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.
\]

**Proof:** Just multiply and verify it works.

\[
\frac{1}{ad - cb} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.
\]

Therefore, this is indeed the inverse.

The expression, \( ad - cb \) is the determinant of the given matrix. Recall, this was discussed in connection with the cross product. This will be discussed in more generality later.

### 5.3 Systems Of Equations And Matrices

Consider a system of linear equations. They all are of this form
\[
\begin{align*}
& a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\
& a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\
& \vdots \\
& a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m
\end{align*}
\]

The above being a system of \( m \) equations in \( n \) variables.

From the way you multiply matrices times vectors, this can be written in the form
\[
\begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \cdots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{pmatrix}
= 
\begin{pmatrix}
b_1 \\
b_2 \\
\vdots \\
b_m
\end{pmatrix}.
\]

Thus, all linear systems of equations can be written in the form
\[
Ax = b
\]

where \( A \) is an \( m \times n \) matrix for \( m, n \) chosen according to the numbers of equations and variables.
Example 5.3.1 Here are some equations. Write in the form $A\mathbf{x} = \mathbf{b}$.

\[
\begin{align*}
    x_1 + 2x_2 + 4x_3 &= -6 \\
    3x_1 + 2x_2 + 5x_4 &= 7 \\
    x_1 - x_2 + x_3 &= 2
\end{align*}
\]

In this case there are three equations and four variables. Thus there should be three rows and four columns in the matrix. You can write it in the desired form as follows.

\[
\begin{pmatrix}
    1 & 2 & 4 & 0 \\
    3 & 2 & 0 & 5 \\
    1 & -1 & 1 & 0 \\
\end{pmatrix}
\begin{pmatrix}
    x_1 \\
    x_2 \\
    x_3 \\
    x_4 \\
\end{pmatrix}
= 
\begin{pmatrix}
    -6 \\
    7 \\
    2 \\
\end{pmatrix}
\]

All other examples are entirely similar. You know how to solve this now. Now here is some terminology.

Definition 5.3.2 Let $A$ be an $m \times n$ matrix. $N(A)$ is called the null space of $A$. It is also denoted as $\ker(A)$. It is defined as follows.

\[
N(A) \equiv \{ \mathbf{x} \in \mathbb{F}^n \text{ such that } A\mathbf{x} = \mathbf{0} \}
\]

In other words, this is just the set of all vectors which $A$ sends to $\mathbf{0}$.

Example 5.3.3 Find $N(A)$ where

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

To do this, you set up the augmented matrix and do row operations till you find the solution. The augmented matrix is

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

Then doing row operations in an auspicious manner, you find the row reduced echelon form is

\[
\begin{pmatrix}
    1 & 0 & 0 & \frac{5}{4} & 0 \\
    0 & 1 & 0 & \frac{5}{8} & 0 \\
    0 & 0 & 1 & -\frac{3}{8} & 0 \\
\end{pmatrix}
\]

Therefore, the null space is of the form. Recall that $x_4$ is called a free variable.

\[
x_3 = \frac{5}{8}x_4, x_2 = -\frac{5}{8}x_4, x_1 = -\frac{5}{4}x_4, x_4 \in \mathbb{F}.
\]

People like to write this in the form

\[
t
\begin{pmatrix}
    -5/4 \\
    -5/8 \\
    5/8 \\
    1 \\
\end{pmatrix}
\]
Of course you could change the parameter by replacing $t$ with $8t$ and this gives the null space equal to

$$
\begin{pmatrix}
-10 \\
-5 \\
5 \\
8 \\
\end{pmatrix},
t \in \mathbb{F}
$$

There is absolutely nothing new here. It is just new jargon applied to the problem already solved.

### 5.4 Elementary Matrices

**Quiz**

1. Here is a matrix.

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

Find its inverse.

2. The inverse of

$$
\begin{pmatrix}
1 & \frac{1}{2} & \frac{1}{2} \\
0 & \frac{1}{2} & \frac{1}{2} \\
1 & -1 & 0 \\
\end{pmatrix}
$$

is

$$
\begin{pmatrix}
1 & -1 & 0 \\
1 & -1 & -1 \\
-1 & 3 & 1 \\
\end{pmatrix}
$$

Use this fact to write the solution to the system

$$
\begin{pmatrix}
x + \frac{1}{2}y + \frac{1}{2}z = a \\
\frac{1}{2}y + \frac{1}{2}z = b \\
x - y = c \\
\end{pmatrix}
$$

in terms of $a, b, c$.

The elementary matrices result from doing a row operation to the identity matrix.

**Definition 5.4.1** The row operations consist of the following

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

The elementary matrices are given in the following definition.

**Definition 5.4.2** The elementary matrices consist of those matrices which result by applying a row operation to an identity matrix. Those which involve switching rows of the identity are called permutation matrices.

---

1 More generally, a permutation matrix is a matrix which comes by permuting the rows of the identity matrix, not just switching two rows.
As an example of why these elementary matrices are interesting, consider the following.

\[
\begin{pmatrix}
0 & 1 & 0 \\
1 & 0 & 0 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
a & b & c & d \\
x & y & z & w \\
f & g & h & i
\end{pmatrix} =
\begin{pmatrix}
a & b & c & d \\
x & y & z & w \\
f & g & h & i
\end{pmatrix}
\]

A 3 × 4 matrix was multiplied on the left by an elementary matrix which was obtained from row operation 1 applied to the identity matrix. This resulted in applying the operation 1 to the given matrix. This is what happens in general.

Now consider what these elementary matrices look like. First consider the one which involves switching row \(i\) and row \(j\) where \(i < j\). This matrix is of the form

\[
\begin{pmatrix}
1 & 0 & & & \\
& \ddots & & & \\
0 & & 1 & & \\
& & \ddots & & \\
1 & & & \ddots & \\
0 & & & & 1
\end{pmatrix}
\]

Note how the \(i^{th}\) and \(j^{th}\) rows are switched in the identity matrix and there are thus all ones on the main diagonal except for those two positions indicated. The two exceptional rows are shown. The \(i^{th}\) row was the \(j^{th}\) and the \(j^{th}\) row was the \(i^{th}\) in the identity matrix.

Now consider what this does to a column vector.

\[
\begin{pmatrix}
1 & 0 & & & \\
& \ddots & & & \\
0 & & 1 & & \\
& & \ddots & & \\
1 & & & \ddots & \\
0 & & & & 1
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_i \\
x_j \\
x_{j} \\
x_{i} \\
x_n
\end{pmatrix} =
\begin{pmatrix}
x_1 \\
x_i \\
x_j \\
x_{j} \\
x_{i} \\
x_n
\end{pmatrix}
\]

Now denote by \(P^{ij}\) the elementary matrix which comes from the identity from switching rows \(i\) and \(j\). From what was just explained consider multiplication on the left by this elementary matrix.

\[
P^{ij}
\begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1p} \\
& & & \\
a_{i1} & a_{i2} & \cdots & a_{ip} \\
& & & \\
a_{j1} & a_{j2} & \cdots & a_{jp} \\
& & & \\
a_{n1} & a_{n2} & \cdots & a_{np}
\end{pmatrix}
\]
From the way you multiply matrices this is a matrix which has the indicated columns.

\[
\begin{pmatrix}
    a_{11} & a_{12} & \cdots & a_{1p} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{i1} & a_{i2} & \cdots & a_{ip} \\
    \vdots & \vdots & \cdots & \vdots \\
    a_{j1} & a_{j2} & \cdots & a_{jp} \\
    \vdots & \vdots & \cdots & \vdots \\
    a_{n1} & a_{n2} & \cdots & a_{np}
\end{pmatrix}
\]

\[
= 
\begin{pmatrix}
   a_{11} & a_{12} & \cdots & a_{1p} \\
   \vdots & \vdots & \ddots & \vdots \\
   a_{i1} & a_{i2} & \cdots & a_{ip} \\
   \vdots & \vdots & \cdots & \vdots \\
   a_{j1} & a_{j2} & \cdots & a_{jp} \\
   \vdots & \vdots & \cdots & \vdots \\
   a_{n1} & a_{n2} & \cdots & a_{np}
\end{pmatrix}
\]
row of the identity matrix is of the form
\[
\begin{pmatrix}
1 & 0 \\
\vdots & \ddots \\
1 & c \\
0 & 1
\end{pmatrix}
\]

Now consider what this does to a column vector.
\[
\begin{pmatrix}
1 & 0 \\
\vdots & \ddots \\
1 & c \\
0 & 1
\end{pmatrix}
\begin{pmatrix}
v_1 \\
v_i-1 \\
v_i \\
v_i+1 \\
v_n
\end{pmatrix}
= \begin{pmatrix}
v_1 \\
v_i-1 \\
cv_i \\
v_i+1 \\
v_n
\end{pmatrix}
\]

Denote by $E(c,i)$ this elementary matrix which multiplies the $i^{th}$ row of the identity by the nonzero constant, $c$. Then from what was just discussed and the way matrices are multiplied,
\[
\begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1p} \\
\vdots & \vdots & & \vdots \\
a_{i1} & a_{i2} & \cdots & a_{ip} \\
\vdots & \vdots & & \vdots \\
a_{j1} & a_{j2} & \cdots & a_{jp} \\
\vdots & \vdots & & \vdots \\
a_{n1} & a_{n2} & \cdots & a_{np}
\end{pmatrix}
\]
equals a matrix having the columns indicated below.
\[
= \begin{pmatrix}
a_{11} \\
\vdots \\
a_{i1} \\
\vdots \\
a_{j1} \\
\vdots \\
a_{n1}
\end{pmatrix}
, E(c,i)
, \begin{pmatrix}
a_{12} \\
\vdots \\
a_{i2} \\
\vdots \\
a_{j2} \\
\vdots \\
a_{n2}
\end{pmatrix}
, \cdots,E(c,i)
, \begin{pmatrix}
a_{1p} \\
\vdots \\
a_{ip} \\
\vdots \\
a_{jp} \\
\vdots \\
a_{np}
\end{pmatrix}
\]
This proves the following lemma.

**Lemma 5.4.5** Let $E(c, i)$ denote the elementary matrix corresponding to the row operation in which the $i^{th}$ row is multiplied by the nonzero constant, $c$. Thus $E(c, i)$ involves multiplying the $i^{th}$ row of the identity matrix by $c$. Then

$$E(c, i)A = B$$

where $B$ is obtained from $A$ by multiplying the $i^{th}$ row of $A$ by $c$.

**Example 5.4.6** Consider this.

$$\begin{pmatrix}
1 & 0 & 0 \\
0 & 5 & 0 \\
0 & 0 & 1
\end{pmatrix} \begin{pmatrix}
a & b \\
c & d \\
e & f
\end{pmatrix} = \begin{pmatrix}
a & b \\
5c & 5d \\
e & f
\end{pmatrix}$$

Finally consider the third of these row operations. Denote by $E(c \times i + j)$ the elementary matrix which replaces the $j^{th}$ row with itself added to $c$ times the $i^{th}$ row added to it. In case $i < j$ this will be of the form

$$\begin{pmatrix}
1 & \cdots & 0 \\
\vdots & 1 & \vdots \\
c & 1 & \vdots \\
0 & \cdots & 0 & 1
\end{pmatrix}$$

Now consider what this does to a column vector.

$$\begin{pmatrix}
1 & \cdots & 0 \\
\vdots & 1 & \vdots \\
c & 1 & \vdots \\
0 & \cdots & 0 & 1
\end{pmatrix} \begin{pmatrix}
v_1 \\
v_j \\
v_j \\
v_n
\end{pmatrix} = \begin{pmatrix}
v_1 \\
v_j \\
(cv_i + v_j) \\
v_n
\end{pmatrix}$$
5.4. ELEMENTARY MATRICES

Now from this and the way matrices are multiplied,

$$E(c \times i + j)$$

$$\begin{pmatrix}
    a_{11} & a_{12} & \cdots & a_{1p} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{i1} & a_{i2} & \cdots & a_{ip} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{j2} & a_{j2} & \cdots & a_{jp} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{n1} & a_{n2} & \cdots & a_{np}
\end{pmatrix}$$

equals a matrix of the following form having the indicated columns.

$$\begin{pmatrix}
    a_{11} \\
    \vdots \\
    a_{i1} \\
    \vdots \\
    a_{j2} \\
    \vdots \\
    a_{n1}
\end{pmatrix}, \begin{pmatrix}
    a_{12} \\
    \vdots \\
    a_{i2} \\
    \vdots \\
    a_{j2} \\
    \vdots \\
    a_{n2}
\end{pmatrix}, \begin{pmatrix}
    a_{1p} \\
    \vdots \\
    a_{ip} \\
    \vdots \\
    a_{jp} \\
    \vdots \\
    a_{np}
\end{pmatrix}$$

The case where $i > j$ is handled similarly. This proves the following lemma.

**Lemma 5.4.7** Let $E(c \times i + j)$ denote the elementary matrix obtained from $I$ by replacing the $j^{th}$ row with $c$ times the $i^{th}$ row added to it. Then

$$E(c \times i + j) A = B$$

where $B$ is obtained from $A$ by replacing the $j^{th}$ row of $A$ with itself added to $c$ times the $i^{th}$ row of $A$.

**Example 5.4.8** Consider the third row operation.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \\ 2a + e & 2b + f \end{pmatrix}$$

The next theorem is the main result.
Theorem 5.4.9 To perform any of the three row operations on a matrix \( A \) it suffices to do the row operation on the identity matrix obtaining an elementary matrix \( E \) and then take the product, \( EA \). Furthermore, each elementary matrix is invertible and its inverse is an elementary matrix.

Proof: The first part of this theorem has been proved in Lemmas 5.4.3 - 5.4.7. It only remains to verify the claim about the inverses. Consider first the elementary matrices corresponding to row operation of type three.

\[
E (-c \times i + j) E (c \times i + j) = I
\]

This follows because the first matrix takes \( c \) times row \( i \) in the identity and adds it to row \( j \). When multiplied on the left by \( E (-c \times i + j) \) it follows from the first part of this theorem that you take the \( i^{th} \) row of \( E (c \times i + j) \) which coincides with the \( i^{th} \) row of \( I \) since that row was not changed, multiply it by \(-c\) and add to the \( j^{th} \) row of \( E (c \times i + j) \) which was the \( j^{th} \) row of \( I \) added to \( c \) times the \( i^{th} \) row of \( I \). Thus \( E (-c \times i + j) \) multiplied on the left, undoes the row operation which resulted in \( E (c \times i + j) \). The same argument applied to the product

\[
E (c \times i + j) E (-c \times i + j)
\]

replacing \( c \) with \(-c\) in the argument yields that this product is also equal to \( I \). Therefore, \( E (c \times i + j)^{-1} = E (-c \times i + j) \).

Similar reasoning shows that for \( E (c, i) \) the elementary matrix which comes from multiplying the \( i^{th} \) row by the nonzero constant, \( c \),

\[
E (c, i)^{-1} = E (c^{-1}, i).
\]

Finally, consider \( P^{ij} \) which involves switching the \( i^{th} \) and the \( j^{th} \) rows.

\[
P^{ij} P^{ij} = I
\]

because by the first part of this theorem, multiplying on the left by \( P^{ij} \) switches the \( i^{th} \) and \( j^{th} \) rows of \( P^{ij} \) which was obtained from switching the \( i^{th} \) and \( j^{th} \) rows of the identity. First you switch them to get \( P^{ij} \) and then you multiply on the left by \( P^{ij} \) which switches these rows again and restores the identity matrix. Thus \( (P^{ij})^{-1} = P^{ij} \).

The geometric significance of these elementary operations is interesting. The following picture shows the effect of doing \( E (\frac{1}{3} \times 3 + 1) \) on a box. You will see that it shears the box in one direction. Of course there would be corresponding shears in the other directions also. Note that this does not change the volume.

The other elementary matrices have similar simple geometric interpretations. For example, \( E (c, i) \) merely multiplies the \( i^{th} \) variable by \( c \). It stretches or contracts the box in
that direction. If $c$ is negative, it also causes the box to be reflected in this direction. The following picture illustrates the effect of $P^{13}$ on a box in three dimensions. It changes the $x$ and the $z$ values.

5.5 Finding Linear Relationships Between Vectors

The row reduced echelon form and more generally the technique of row operations, are tools for discovering hidden linear relationships between column vectors. You have already been doing this. Consider a system of equations, for example

\[
\begin{align*}
    x + 2y - z &= 3 \\
    2x + y + z &= 4 \\
    x - y + 3z &= 7 \\
\end{align*}
\]

What are you really asking for when you request the solution? Isn’t it to find scalars $x, y, z$ such that

\[
x \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + y \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix} + z \begin{pmatrix} -1 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 7 \end{pmatrix}
\]

In other words, you are attempting to find the vector on the right as a linear combination of the three vectors on the left. That is what the expression on the left is called, a linear combination of the three vectors which occur there. When the scalars $x, y, z$ are found, the result is a linear relationship between these vectors. For the sake of consistency, we like to write it as

\[
x \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + y \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix} + z \begin{pmatrix} -1 \\ 1 \\ 3 \end{pmatrix} + (-1) \begin{pmatrix} 3 \\ 4 \\ 7 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
\]

but this is of no importance. A linear relation or linear relationship is just an equation involving a linear combination of vectors. The problem of solving a linear system of equations is just the process of finding a linear relationship like the above. In this case, $x = -13/3, y = 20/3, z = 6$, not the first thing you would think of. In other words, the desired linear relationship was hidden.

Recall how you found the answer. You formed the matrix

\[
\begin{pmatrix} 1 & 2 & -1 & 3 \\ 2 & 1 & 1 & 4 \\ 1 & -1 & 3 & 7 \end{pmatrix}
\]
and then you did row operations till it was easy to spot the desired solution (linear relationship) between the columns. The more general problem is to find linear relationships between the columns of an arbitrary matrix in a systematic way.

The following definition is the precise description of what is meant by a linear combination.

**Definition 5.5.1** The vector, $\mathbf{u}$ is a **linear combination** of the vectors, $\mathbf{v}_1, \cdots, \mathbf{v}_m$ if there exist scalars, $c_1, \cdots, c_m$ such that

$$
\mathbf{u} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + \cdots + c_m \mathbf{v}_m = \sum_{k=1}^{m} c_k \mathbf{v}_k.
$$

A **linear relationship** between the vectors $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is an expression of the form

$$
0 = \sum_{k=1}^{n} c_k \mathbf{v}_k
$$

where the $c_1, \cdots, c_n$ are scalars or more generally, any equation involving a sum of scalars times vectors. When some of the scalars are nonzero, it is also called a **dependence relation**.

**Example 5.5.2**

$$
3 \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix} + 5 \begin{pmatrix} 1 \\ -2 \\ 4 \end{pmatrix} + (-2) \begin{pmatrix} 5 \\ 3 \\ 7 \end{pmatrix} = \begin{pmatrix} 1 \\ -7 \\ 9 \end{pmatrix}
$$

Thus $\begin{pmatrix} 1 \\ -7 \\ 9 \end{pmatrix}$ is a linear combination of the vectors, $\begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 1 \\ -2 \\ 4 \end{pmatrix}$, and $\begin{pmatrix} 5 \\ 3 \\ 7 \end{pmatrix}$. In this case the scalars are 3, 5, and $-2$. A linear relationship between these four vectors is

$$
3 \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix} + 5 \begin{pmatrix} 1 \\ -2 \\ 4 \end{pmatrix} + (-2) \begin{pmatrix} 5 \\ 3 \\ 7 \end{pmatrix} + (-1) \begin{pmatrix} 1 \\ -7 \\ 9 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
$$

**Example 5.5.3** Consider the vectors $\begin{pmatrix} 6 \\ 6 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 0 \\ 12 \\ 8 \end{pmatrix}$, $\begin{pmatrix} 2 \\ 6 \\ 3 \end{pmatrix}$. Then

$$
2 \begin{pmatrix} 6 \\ 6 \\ 0 \end{pmatrix} + 3 \begin{pmatrix} 0 \\ 12 \\ 8 \end{pmatrix} + (-6) \begin{pmatrix} 2 \\ 6 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
$$

Check it and you will see I am right.

How did I know what scalars would produce the above linear relationship? This involves the fundamental lemma on row operations which says that row operations done on a matrix preserve all linear relationships involving the columns.
5.5. FINDING LINEAR RELATIONSHIPS BETWEEN VECTORS

5.5.1 The Great And Glorious Lemma On Row Operations

First recall the row operations.

Definition 5.5.4 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.

Here is the great and glorious lemma on row operations. This lemma says that when you do a row operations, all linear relationships between the columns are preserved. Because of this, you can discover linear relationships between the columns by doing row operations till you get a matrix, usually the row reduced echelon form, in which all these linear relationships are completely obvious. Then by this lemma, about to be presented, the same linear relationships involving the columns exist for the original matrix.

Lemma 5.5.5 Let $B$ and $A$ 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: Let $A$ equal the following matrix in which the $a_k$ are the columns

$$
\begin{pmatrix}
a_1 & a_2 & \cdots & a_n
\end{pmatrix}
$$

and let $B$ equal the following matrix in which the columns are given by the $b_k$

$$
\begin{pmatrix}
b_1 & b_2 & \cdots & b_n
\end{pmatrix}
$$

Then by Theorem 5.4.9 on Page 128 $b_k = E a_k$ where $E$ is an elementary matrix. Suppose then that one of the columns of $A$ is a linear combination of some other columns of $A$. Say

$$
a_k = \sum_{r \in S} c_r a_r.
$$

Then multiplying by $E$,

$$
b_k = E a_k = \sum_{r \in S} c_r E a_r = \sum_{r \in S} c_r b_r.
$$

This proves the lemma.

Example 5.5.6 Discover linear relationships involving the columns of the matrix

$$
\begin{pmatrix}
1 & 2 & 1 & 5 & 2 & 7 \\
1 & 2 & -1 & 1 & 0 & -1 \\
1 & 2 & 1 & 5 & 1 & 4
\end{pmatrix}
$$
First take \((-1)\) times the top row and add to the second and then \((-1)\) times the top row and add to the third. This yields

\[
\begin{pmatrix}
1 & 2 & 1 & 5 & 2 & 7 \\
0 & 0 & -2 & -4 & -2 & -8 \\
0 & 0 & 0 & 0 & -1 & -3
\end{pmatrix}
\]

Next take \((1/2)\) times the middle row and add it to the top and then multiply the middle row by \(-1/2\).

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

Finally add the bottom row to the middle row and then add the bottom row to the top and then multiply the bottom row by \(-1\).

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

(5.22)

Note the result is in row reduced echelon form. According to the great and glorious lemma each row operation preserved all linear relationships among the columns. The nice thing about (5.22) is that these linear relationships are all obvious. Thus the second column is 2 times the first. Of course this is obvious in the original matrix also. Now consider the fourth column.

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

That is the fourth column equals three times the first plus 2 times the third. By the great and glorious lemma, the same will be true of the columns of the original matrix. Thus

\[
\begin{pmatrix}
5 \\
1 \\
5
\end{pmatrix}
= 3 \begin{pmatrix}
1 \\
1 \\
1
\end{pmatrix}
+ 2 \begin{pmatrix}
1 \\
-1 \\
1
\end{pmatrix}
\]

The last column in (5.22) equals the third column plus 3 times the fifth. Therefore, the same must be true of the corresponding columns in the original matrix. Thus

\[
\begin{pmatrix}
7 \\
-1 \\
4
\end{pmatrix}
= \begin{pmatrix}
1 \\
-1 \\
1
\end{pmatrix}
+ 3 \begin{pmatrix}
2 \\
0 \\
1
\end{pmatrix}
\]

I will leave it to you to spot other linear relationships among the columns by using (5.22), the row reduced echelon form of the original matrix. The point is that they are all completely transparent when you are looking at the row reduced echelon form although they may be hidden in the original matrix.

A linear relationship which exhibits one vector as a linear combination of some others is also called a dependence relation.
5.5.2 Theory Of Row Reduced Echelon Form

Note the use of the definite article, “the” in referring to row reduced echelon form. A natural question is whether a given matrix could have more than one row reduced echelon form. The answer which I will show here is no. There is only one. The proof of this astonishing result also depends on the great and glorious lemma. If you are willing to believe this astonishing result, you don’t need to read the rest of this section.

To begin with I will present a review of the row reduced echelon form in slightly different terms than those used earlier. It is convenient to describe it slightly differently to use Lemma 5.5.5.

**Definition 5.5.7** 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 $e_1, e_2, \cdots, e_k$, the next column is either $e_{k+1}$ or is a linear combination of the vectors, $e_1, e_2, \cdots, e_k$.

There are two aspects of the row reduced echelon form which are important. One is its existence and the other is its uniqueness. The following theorem deals with its existence.

**Theorem 5.5.8** 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 $R^{m-1}$. 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,

$$
\begin{pmatrix}
* & * & * & * \\
\vdots & \vdots & \vdots & \vdots \\
* & * & * & *
\end{pmatrix}
$$
and if the first column consists of all zeros but the next one is not all zeros, you would get something like this.

\[
\begin{pmatrix}
0 & 1 & * & * & * \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & * & * & * \\
\end{pmatrix}
\]

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.

\[
\begin{pmatrix}
0 & 1 & 0 & * & * & * \\
0 & 0 & 0 & 1 & * & * \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & * & * \\
\end{pmatrix}
\]

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.

Now here is some terminology.

**Definition 5.5.9** 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 theorem says this is the case.

**Definition 5.5.10** Two matrices are said to be **row equivalent** if one can be obtained from the other by a sequence of row operations.

It has been shown above that every matrix is row equivalent to one which is in row reduced echelon form.

**Theorem 5.5.11** The row reduced echelon form is unique. That is if \( B, C \) are two matrices in row reduced echelon form and both are row equivalent to \( A \), then \( B = C \).

**Proof:** Suppose \( B \) and \( C \) are both row reduced echelon forms for the matrix, \( A \). Then they clearly have the same zero columns since row operations leave zero columns unchanged. If \( B \) has the sequence \( e_1, e_2, \ldots, e_r \) occurring for the first time in the positions, \( i_1, i_2, \ldots, i_r \), the description of the row reduced echelon form means that each of these columns is not a linear combination of the preceding columns. Therefore, by Lemma 5.5.5, the same is true of the columns in positions \( i_1, i_2, \ldots, i_r \) for \( C \). It follows from the description of the row reduced echelon form that \( e_1, \ldots, e_r \) occur respectively for the first time in columns \( i_1, i_2, \ldots, i_r \) for \( C \). Therefore, both \( B \) and \( C \) have the sequence \( e_1, e_2, \ldots, e_r \) occurring for the first time in the positions, \( i_1, i_2, \ldots, i_r \). By Lemma 5.5.5, the columns between the \( i_k \) and \( i_{k+1} \) position in the two matrices are linear combinations involving the same scalars of the columns in the \( i_1, \ldots, i_k \) position. Also the columns after the \( i_r \) position are linear combinations of the columns in the \( i_1, \ldots, i_r \) positions involving the same scalars in both matrices. This is equivalent to the assertion that each of these columns is identical and this proves the corollary.

Here is another very important proposition which has to do with whether \( A \) is one to one.
5.5. FINDING LINEAR RELATIONSHIPS BETWEEN VECTORS

Proposition 5.5.12 Suppose $A$ is an $m \times n$ matrix where $n > m$. Then $A$ is not one to one. That is there exists $x \neq 0$ such that $Ax = 0$.

Proof: Since there are more columns than rows, it follows the row reduced echelon form cannot consist entirely of vectors $e_k$. Thus there must be a free variable and so there exists $x \neq 0$ such that $Ax = 0$.

Now here is another version of this.

Corollary 5.5.13 Suppose $A$ is an $n \times n$ matrix. Then $A$ is one to one if and only if its row reduced echelon form is the identity matrix.

Proof: Form the augmented matrix

\[
\begin{pmatrix}
A & 0 \\
R & 0
\end{pmatrix}
\]

where $R$ is the row reduced echelon form of $A$. $A$ is not one to one if and only if there exists at least one free variable. Therefore, there is a free variable and $A$ is not one to one if and only if $R \neq I$. Thus $A$ is one to one if and only if $R = I$. This proves the corollary.

The following corollary follows.

Corollary 5.5.14 Let $A$ be an $m \times n$ matrix and let $R$ denote the row reduced echelon form obtained from $A$ by row operations. Then there exists a sequence of elementary matrices, $E_1, \cdots, E_p$ such that

\[(E_pE_{p-1} \cdots E_1) A = R.\]

Proof: This follows from the fact that row operations are equivalent to multiplication on the left by an elementary matrix.

Corollary 5.5.15 Let $A$ be an $n \times n$ matrix. Then $A$ is invertible if and only if $A$ equals a finite product of elementary matrices.

Proof: Suppose first that $A^{-1}$ is given to exist. It follows $A$ must be one to one because if $Ax = 0$, then doing $A^{-1}$ to both sides yields $x = Ix = A^{-1}Ax = A^{-1}0 = 0$.

By Corollary 5.5.13 the row reduced echelon form of $A$ is $I$ and so by Corollary 5.5.14, there is a sequence of elementary matrices, $E_1, \cdots, E_p$ such that

\[(E_pE_{p-1} \cdots E_1) A = I.\]

But now multiply on the left on both sides by $E_p^{-1}$ then by $E_{p-1}^{-1}$ and then by $E_{p-2}^{-1}$ etc. until you get

\[A = E_1^{-1}E_2^{-1} \cdots E_{p-1}^{-1}E_p^{-1}\]

and by Theorem 5.4.14 each of these in this product is an elementary matrix.

Next suppose $A = E_1E_2 \cdots E_m$ where each $E_k$ is an elementary matrix. Since each of these is invertible, it follows so is $A$ and in fact

\[A^{-1} = E_m^{-1}E_{m-1}^{-1} \cdots E_1^{-1},\]

the product of the inverses in the reverse order. This proves the corollary.
5.6 Block Multiplication Of Matrices

Consider the following problem

\[
\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix}
\]

You know how to do this. You get

\[
\begin{pmatrix} AE + BG & AF + BH \\ CE + DG & CF + DH \end{pmatrix}.
\]

Now what if instead of numbers, the entries, \(A, B, C, D, E, F, G\) are matrices of a size such that the multiplications and additions needed in the above formula all make sense. Would the formula be true in this case? I will show below that this is true.

Suppose \(A\) is a matrix of the form

\[
A = \begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{r1} & \cdots & A_{rm} \end{pmatrix}
\]

(5.23)

where \(A_{ij}\) is a \(s_i \times p_j\) matrix where \(s_i\) is constant for \(j = 1, \ldots, m\) for each \(i = 1, \ldots, r\). Such a matrix is called a block matrix, also a partitioned matrix. How do you get the block \(A_{ij}\)? Here is how for \(A\) an \(m \times n\) matrix:

\[
\begin{pmatrix} \begin{pmatrix} 0 & I_{s_i \times s_i} & 0 \\ \vdots \end{pmatrix} & A \begin{pmatrix} 0 \\ I_{p_j \times p_j} \\ 0 \end{pmatrix} \end{pmatrix}.
\]

(5.24)

In the block column matrix on the right, you need to have \(c_j - 1\) rows of zeros above the small \(p_j \times p_j\) identity matrix where the columns of \(A\) involved in \(A_{ij}\) are \(c_j, \ldots, c_j + p_j\) and in the block row matrix on the left, you need to have \(r_i - 1\) columns of zeros to the left of the \(s_i \times s_i\) identity matrix where the rows of \(A\) involved in \(A_{ij}\) are \(r_i, \ldots, r_i + s_i\). An important observation to make is that the matrix on the right specifies columns to use in the block and the one on the left specifies the rows used. There is no overlap between the blocks of \(A\). Thus the identity \(n \times n\) identity matrix corresponding to multiplication on the right of \(A\) is of the form

\[
\begin{pmatrix} I_{p_1 \times p_1} & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & I_{p_m \times p_m} \end{pmatrix}
\]

these little identity matrices don’t overlap. A similar conclusion follows from consideration of the matrices \(I_{s_i \times s_i}\).

Next consider the question of multiplication of two block matrices. Let \(B\) be a block matrix of the form

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

(5.25)
and $A$ is a block matrix of the form

$$
\begin{pmatrix}
A_{11} & \cdots & A_{1m} \\
\vdots & \ddots & \vdots \\
A_{p1} & \cdots & A_{pm}
\end{pmatrix}
$$

and that for all $i, j$, it makes sense to multiply $B_{is}A_{sj}$ for all $s \in \{1, \cdots, p\}$. (That is the two matrices, $B_{is}$ and $A_{sj}$ are conformable.) and that for fixed $ij$, it follows $B_{is}A_{sj}$ is the same size for each $s$ so that it makes sense to write $\sum_s B_{is}A_{sj}$.

The following theorem says essentially that when you take the product of two matrices, you can do it two ways. One way is to simply multiply them forming $BA$. The other way is to partition both matrices, formally multiply the blocks to get another block matrix and this one will be $BA$ partitioned. Before presenting this theorem, here is a simple lemma which is really a special case of the theorem.

**Lemma 5.6.1** Consider the following product.

$$
\begin{pmatrix}
0 \\
I \\
0
\end{pmatrix}
\begin{pmatrix}
0 & I & 0 \\
0 & I & 0 \\
0 & I & 0
\end{pmatrix}
$$

where the first is $n \times r$ and the second is $r \times n$. The small identity matrix $I$ is an $r \times r$ matrix and there are $l$ zero rows above $I$ and $l$ zero columns to the left of $I$ in the right matrix. Then the product of these matrices is a block matrix of the form

$$
\begin{pmatrix}
0 & 0 & 0 \\
0 & I & 0 \\
0 & 0 & 0
\end{pmatrix}
$$

**Proof:** From the definition of the way you multiply matrices, the product is

$$
\begin{pmatrix}
0 \\
I \\
0
\end{pmatrix}
\begin{pmatrix}
0 & \cdots & 0 \\
0 & \cdots & 0 \\
0 & \cdots & 0
\end{pmatrix}
\begin{pmatrix}
e_1 \\
e_r \\
e_j
\end{pmatrix}
\begin{pmatrix}
0 & \cdots & 0 \\
0 & \cdots & 0 \\
0 & \cdots & 0
\end{pmatrix}
\begin{pmatrix}
0 \\
I \\
0
\end{pmatrix}
\begin{pmatrix}
0 & \cdots & 0 \\
0 & \cdots & 0 \\
0 & \cdots & 0
\end{pmatrix}
$$

which yields the claimed result. In the formula $e_j$ refers to the column vector of length $r$ which has a 1 in the $j^{th}$ position. This proves the lemma.

**Theorem 5.6.2** Let $B$ be a $q \times p$ block matrix as in (5.25) and let $A$ be a $p \times n$ block matrix as in (5.26) such that $B_{is}$ is conformable with $A_{sj}$ and each product, $B_{is}A_{sj}$ for $s = 1, \cdots, p$ is of the same size so they can be added. Then $BA$ can be obtained as a block matrix such that the $ij^{th}$ block is of the form

$$
\sum_s B_{is}A_{sj}.
$$

**Proof:** From (5.26)

$$
B_{is}A_{sj} = \begin{pmatrix}
0 & I_{r_i \times r_i} & 0 \\
0 & I_{p_s \times p_s} & 0 \\
0 & 0 & I_{q_j \times q_j}
\end{pmatrix}.
$$

$$
\begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}
\begin{pmatrix}
0 & I_{p_s \times p_s} & 0 \\
0 & I_{q_j \times q_j}
\end{pmatrix}
$$
CHAPTER 5. MATRICES

where here it is assumed \( B_{is} \) is \( r_i \times p_s \) and \( A_{sj} \) is \( p_s \times q_j \). The product involves the \( s^{th} \) block in the \( i^{th} \) row of blocks for \( B \) and the \( s^{th} \) block in the \( j^{th} \) column of \( A \). Thus there are the same number of rows above the \( I_{ps \times ps} \) as there are columns to the left of \( I_{ps \times ps} \) in those two inside matrices. Then from Lemma 5.6.1

\[
\begin{pmatrix}
0 \\
I_{ps \times ps} \\
0
\end{pmatrix}
\begin{pmatrix}
0 & I_{ps \times ps} \\
0 & 0
\end{pmatrix} =
\begin{pmatrix}
0 & 0 & 0 \\
0 & I_{ps \times ps} \\
0 & 0 & 0
\end{pmatrix}
\]

Since the blocks of small identity matrices do not overlap,

\[
\sum_s \begin{pmatrix}
0 & 0 & 0 \\
0 & I_{ps \times ps} \\
0 & 0 & 0
\end{pmatrix} =
\begin{pmatrix}
I_{p_1 \times p_1} & 0 \\
\vdots & \ddots \\
0 & 0 & I_{ps \times ps}
\end{pmatrix} = I
\]

and so

\[
\sum_s B_{is}A_{sj} =
\]

\[
\sum_s \begin{pmatrix}
0 & I_{r_i \times r_i} & 0 \\
0 & 0 & 0
\end{pmatrix} B \begin{pmatrix}
0 & I_{ps \times ps} \\
0 & 0
\end{pmatrix} A \begin{pmatrix}
0 & I_{q_j \times q_j} \\
0 & 0
\end{pmatrix} =
\begin{pmatrix}
0 & I_{r_i \times r_i} \\
I_{q_j \times q_j} & 0
\end{pmatrix} BA
\begin{pmatrix}
0 \\
I_{q_j \times q_j}
\end{pmatrix}
\]

Hence the \( ij^{th} \) block of \( BA \) equals the formal multiplication according to matrix multiplication,

\[
\sum_s B_{is}A_{sj}.
\]

This proves the theorem.

**Example 5.6.3** Let an \( n \times n \) matrix have the form

\[
A = \begin{pmatrix}
a & b \\
c & P
\end{pmatrix}
\]

where \( P \) is \( n - 1 \times n - 1 \). Multiply it by

\[
B = \begin{pmatrix}
p & q \\
r & Q
\end{pmatrix}
\]

where \( B \) is also an \( n \times n \) matrix and \( Q \) is \( n - 1 \times n - 1 \).

You use block multiplication

\[
\begin{pmatrix}
a & b \\
c & P
\end{pmatrix}
\begin{pmatrix}
p & q \\
r & Q
\end{pmatrix} =
\begin{pmatrix}
ap + br & aq + bQ \\
pc + Pr & cq + PQ
\end{pmatrix}
\]

Note that this all makes sense. For example, \( b = 1 \times n - 1 \) and \( r = n - 1 \times 1 \) so \( br \) is a \( 1 \times 1 \). Similar considerations apply to the other blocks.
5.7 Exercises With Answers

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.

\[-3A = -3 \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 7 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -3 & -6 & -9 \\ -6 & -9 & -21 \\ -3 & 0 & -3 \end{pmatrix}, \]

\[3B - A \text{ is nonsense because the matrices } B \text{ and } A \text{ are not of the same size.} \]

\[AC = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 7 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 10 & 7 \\ 18 & 14 \\ 2 & 3 \end{pmatrix}.\]

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 \( x^T y \).

\[
x^T y = \begin{pmatrix} -1 \\ 0 \\ 3 \end{pmatrix} \begin{pmatrix} 3 & 1 & 2 \\ 0 & 0 & 0 \\ 9 & 3 & 6 \end{pmatrix} = \begin{pmatrix} -3 & -1 & -2 \\ 0 & 0 & 0 \\ 9 & 3 & 6 \end{pmatrix}.
\]

4. Write

\[
\begin{pmatrix} 4x_1 - x_2 + 2x_3 \\ 2x_3 + 7x_1 \\ 2x_3 \\ 3x_3 + 3x_2 + x_1 \end{pmatrix}
\]

in the form \( Ax \) 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.

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

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.

\[
\begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 5 & 2 & 0 \\ 2 & 1 & -3 & 2 \\ 1 & 2 & 1 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} -3 & \frac{1}{6} & \frac{5}{6} & \frac{13}{6} \\ 1 & \frac{1}{5} & -\frac{1}{5} & -\frac{5}{6} \\ -1 & 0 & 0 & 1 \\ 1 & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \end{pmatrix}.
\]

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, \( A^{-1} = A^{-1} (AB) = (A^{-1} A) B = IB = B \).
8. Suppose $A, B$ are two matrices. Show $(AB)^{-1} = B^{-1}A^{-1}$.

All you have to do is multiply it. If it acts like the inverse, it is the inverse.

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

Therefore, $B^{-1}A^{-1} = (AB)^{-1}$.

9. Show $(A^{-1})^T = (A^T)^{-1}$.

$$A^T(A^{-1})^T = (A^{-1}A)^T = I^T = I$$ and so $(A^{-1})^T = (A^T)^{-1}$.

10. Here are elementary matrices. Find their inverses.

   (a) \[
   \begin{pmatrix}
   1 & 0 & 0 \\
   a & 1 & 0 \\
   0 & 0 & 1
   \end{pmatrix}
   \begin{pmatrix}
   1 & 0 & 0 \\
   a & 1 & 0 \\
   0 & 0 & 1
   \end{pmatrix}^{-1} = \begin{pmatrix}
   1 & 0 & 0 \\
   -a & 1 & 0 \\
   0 & 0 & 1
   \end{pmatrix}
   \]

   (b) \[
   \begin{pmatrix}
   1 & 0 & 0 \\
   0 & 2 & 0 \\
   0 & 0 & 1
   \end{pmatrix}
   \begin{pmatrix}
   1 & 0 & 0 \\
   0 & 2 & 0 \\
   0 & 0 & 1
   \end{pmatrix}^{-1} = \begin{pmatrix}
   1 & 0 & 0 \\
   0 & \frac{1}{2} & 0 \\
   0 & 0 & 1
   \end{pmatrix}
   \]

   (c) \[
   \begin{pmatrix}
   0 & 1 & 0 \\
   1 & 0 & 0 \\
   0 & 0 & 1
   \end{pmatrix}
   \begin{pmatrix}
   0 & 1 & 0 \\
   1 & 0 & 0 \\
   0 & 0 & 1
   \end{pmatrix}^{-1} = \begin{pmatrix}
   0 & 1 & 0 \\
   1 & 0 & 0 \\
   0 & 0 & 1
   \end{pmatrix}
   \]

11. When you have an $n \times n$ matrix, $A, A^n = A \times A \times \cdots \times A$. If $A^{-1}$ exists, show $(A^{-1})^n = (A^n)^{-1}$.

   The equation is true if $n = 1$. Suppose it is true for $n$. Then by the induction hypothesis,
   \[
   (A^{-1})^{n+1} = (A^{-1})^n A^{-1} = (A^n)^{-1} A^{-1} = (A(A^n))^{-1} = (A^{n+1})^{-1}.
   \]

5.8 The Rank Of A Matrix

Corresponding to such a rectangular array of numbers, there is a row reduced echelon form discussed above. Since the row reduced echelon form is unique by Theorem 5.5.11 it follows it is possible to define things in terms of this unique row reduced echelon form. The rank of a matrix is defined as follows.
**Definition 5.8.1** The rank of a matrix, $A$, equals the number of nonzero rows in its row reduced echelon form. This is the same as the number of pivot columns.

**Example 5.8.2** Find the rank of the matrix,

$$ A = \begin{pmatrix} 1 & 2 & 3 & 1 \\ 0 & 2 & 1 & 1 \\ 1 & 4 & 4 & 2 \end{pmatrix} $$

To find the rank, you obtain the row reduced echelon form and count the number of nonzero rows or equivalently the number of pivot columns. First take $-1$ times the top row and add to the bottom row. This yields

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

Now add $-1$ times the second row to the bottom. This yields

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

Now take $-1$ times the second row and add to the top.

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

Finally, multiply the second row by $1/2$ to get

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

which is in row reduced echelon form. The rank of this matrix is therefore 2.

Note that from the process used to obtain the row reduced echelon form, once you have obtained an echelon form, you know the correct number of rows in the final result. Thus you can simply take the number of nonzero rows in an echelon form and this will be the rank. Note also that the rank is the number of pivot columns. In this case the pivot columns are the first two.

**Definition 5.8.3** A homogeneous system of linear equations is one with augmented matrix of the form

$$ \begin{pmatrix} A & 0 \end{pmatrix} $$

where 0 is a column of zeros and $A$ is an $m \times n$ matrix.

**Example 5.8.4** An example of a homogeneous system of equations is $x + y = 0, 3x - y = 0$. It has augmented matrix,

$$ \begin{pmatrix} 1 & 1 & 0 \\ 3 & -1 & 0 \end{pmatrix} $$
The nice thing about homogeneous systems is that they are always consistent. Simply let all the variables equal zero and you obtain a solution. However, there may be other solutions besides this one. This is related to the concept of rank and free variables.

**Theorem 5.8.5** Let \( A \) be an \( m \times n \) matrix. Form the augmented matrix,

\[
\left( \begin{array}{c|c} A & 0 \end{array} \right)
\]

where 0 is the column of zeros. Thus \( A \) is the coefficient matrix of a system of linear equations with \( n \) variables. Then the number of free variables \( = n - \text{rank} \, (A) \).

**Proof:** The basic variables correspond to the pivot columns of \( A \) and the free variables correspond to the other columns. However, the rank of \( A \) equals the number of pivot columns.

As a corollary here is a theorem which is called the Rank theorem.

**Corollary 5.8.6** (Rank Theorem) Let \( A \) be an \( m \times n \) matrix. Form the augmented matrix,

\[
\left( \begin{array}{c|c} A & b \end{array} \right)
\]

where \( b \) is an \( m \times 1 \) column. Thus \( A \) is the coefficient matrix of a system of linear equations with \( n \) variables. Then if the system of equations represented by the above augmented matrix is consistent, number of free variables \( = n - \text{rank} \, (A) \).

**Proof:** Since the equations represented by the above augmented matrix are consistent, the same argument as in Theorem 5.8.5 holds. The leading entry in the last nonzero row cannot be in the last column because if it were, then the system would fail to be consistent.

### 5.9 Exercises With Answers

1. Find the distance from the point, \((1, 2, 1)\) to the plane \(3x + y - z = 7\).

   You can use the stupid formula for this.

   \[
   \frac{|3 + 2 - 1 - 7|}{\sqrt{9 + 1 + 1}} = \frac{3}{\sqrt{11}}
   \]

2. Find the cosine of the angle between the planes \(x - y + z = 7\) and \(2x + y - 3z = 4\).

   You just need to consider the normal vectors which are \((1, -1, 1)\) and \((2, 1, -3)\). Then the cosine of the angle desired is

   \[
   \cos \theta = \left| \frac{(2, 1, -3) \cdot (1, -1, 1)}{\sqrt{1 + 1 + 1} \sqrt{4 + 1 + 9}} \right| = \frac{1}{2} \sqrt{3/14}
   \]

3. Here are vector equations for two lines. \((x, y, z) = (1, 2, 0) + t (2, 1, 1)\) and \((x, y, z) = (3, 0, 1) + t (1, -2, 1)\). The angle between the direction vectors is not 0 or \(\pi\) and so the lines are not parallel. If they were two lines in \(\mathbb{R}^2\), this means they would need to intersect. However, these two lines do not intersect. If they did, there would exist \(s, t\) such that

   \[(1, 2, 0) + t (2, 1, 1) = (3, 0, 1) + s (1, -2, 1)\]
and this would require the following system of equations would need to hold.

\[
\begin{align*}
1 + 2t &= 3 + s \\
2 + t &= -2s \\
t &= 1 + s
\end{align*}
\]

The augmented matrix for this system is

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

The row reduced echelon form is

\[
\begin{pmatrix}
1 & 0 & | & 0 \\
0 & 1 & | & 0 \\
0 & 0 & | & 1
\end{pmatrix}
\]

and so there is no solution. These lines are called \textbf{skew lines}. Imagine two airplanes, one going from South to North and the other going from East to West. The first travels at 40000 feet and the second at 35000 feet. Their paths never cross. Of course the extra dimension is not present in two dimensions and so their paths would cross if they were moving in a plane. Note also that to consider the question whether the lines intersect, you must look at possibly different values for the parameters.

4. Let two skew lines be given in Problem 3. Find two parallel planes which contain the two lines.

This is easy if you can find the normal vector of the two planes. To say the planes are parallel requires them to have the same normal vector. The two lines were \((x, y, z) = (1, 2, 0) + t(2, 1, 1)\) and \((x, y, z) = (3, 0, 1) + t(1, -2, 1)\). Therefore, the normal vector needs to be perpendicular to both direction vectors. You need \(\mathbf{n} = (2, 1, 1) \times (1, -2, 1) = (3, -1, -5)\). Now the equation of the first plane is

\[(3, -1, -5) \cdot (x - 1, y - 2, z) = 0\]

and the equation of the second plane is

\[(3, -1, -5) \cdot (x - 3, y, z - 1) = 0\]

The two planes are therefore, \(3x - y - 5z = 1\) and \(3x - y - 5z = 4\). You see these are parallel planes because they have the same normal vector and the first contains the first line while the second contains the second line.

5. 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 & \[ & * & | & \[ \\
0 & 0 & 0 & 0 & | & * \\
\end{pmatrix}
\]
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.

6. Here is an augmented matrix in which $\ast$ denotes an arbitrary number and $\blacksquare$ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

$$
\begin{pmatrix}
\blacksquare & \ast & \ast & | & \ast \\
0 & 0 & \blacksquare & | & \blacksquare \\
0 & 0 & \ast & | & 0
\end{pmatrix}
$$

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:

$$
\begin{pmatrix}
\blacksquare & \ast & \ast & | & \ast \\
0 & 0 & \blacksquare & | & \blacksquare \\
0 & 0 & 0 & | & \blacksquare
\end{pmatrix}
$$

This would give a row of zeros equal to something nonzero.

7. 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, $h = 7/3$.

8. Determine if the system is consistent.

$$
x + 2y + 3z - w = 2 \\
x - y + 2z + w = 1 \\
2x + 3y - z = 1 \\
4x + 2y + z = 5
$$

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.

9. 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 \(y = -11/17\).

10. 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.

11. 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
\[x = 2 - 8t, y = 4t, z = t.\]
12. Find the point, \((x_1, y_1)\) which lies on both lines, \(x + 2y = 1\) and \(3x - y = 3\).

   The solution is \(y = 0\) and \(x = 1\).

13. Find the point of intersection of the two lines \(x + y = 3\) and \(x + 2y = 1\).

   The solution is \((5, -2)\).

14. 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.

   To solve this set up the augmented matrix and go to work on it. The augmented matrix is

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

   A reduced echelon matrix for this is

   \[
   \begin{pmatrix}
   1 & 0 & \frac{3}{5} \\
   0 & 1 & \frac{1}{5} \\
   0 & 0 & 0
   \end{pmatrix}
   \]

   Therefore, there is a point in the intersection of these and it is \(y = 1/5\) and \(x = 3/5\). Thus the point is \((3/5, 1/5)\).

15. Do the three planes, \(x + 2y - 3z = 2\), \(x + y + z = 1\), 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.

   You need to find \((x, y, z)\) which solves each equation. The augmented matrix is

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

   A reduced echelon form for the matrix is

   \[
   \begin{pmatrix}
   1 & 0 & 0 & -2 \\
   0 & 1 & 0 & \frac{13}{5} \\
   0 & 0 & 1 & \frac{2}{5}
   \end{pmatrix}
   \]

   and so you should let \((x, y, z) = (-2, 13/5, 2/5)\).

16. Here is an augmented matrix in which \(*\) denotes an arbitrary number and \(\blacksquare\) denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

   \[
   \begin{pmatrix}
   \blacksquare & * & * & * & * & * \\
   0 & \blacksquare & * & * & 0 & * \\
   0 & 0 & \blacksquare & * & * & * \\
   0 & 0 & 0 & \blacksquare & | & *
   \end{pmatrix}
   \]
You could do another set of row operations and reduce the matrix to one of the form

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

It follows there exists a solution but the solution is not unique because \(x_4\) is a free variable. You can pick it to be anything you like and the system will yield values for the other variables.

17. Here is an augmented matrix in which \(\ast\) denotes an arbitrary number and \(\blacksquare\) denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

\[
\begin{pmatrix}
\blacksquare & \ast & \ast & | & \ast \\
0 & \blacksquare & \ast & | & \ast \\
0 & 0 & \blacksquare & | & \ast \\
\end{pmatrix}
\]

In this case there is a unique solution to the system. To see this, you could do more row operations and reduce this to something of the form

\[
\begin{pmatrix}
\blacksquare & 0 & 0 & | & \ast \\
0 & \blacksquare & 0 & | & \ast \\
0 & 0 & \blacksquare & | & \ast \\
\end{pmatrix}
\]

18. Here is an augmented matrix in which \(\ast\) denotes an arbitrary number and \(\blacksquare\) denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

\[
\begin{pmatrix}
\blacksquare & \ast & \ast & \ast & | & \ast \\
0 & \blacksquare & 0 & \ast & 0 & | & \ast \\
0 & 0 & \blacksquare & \ast & 0 & | & \ast \\
0 & 0 & 0 & \blacksquare & | & \ast \\
\end{pmatrix}
\]

In this case, you could do more row operations and get something of the form

\[
\begin{pmatrix}
\blacksquare & 0 & \ast & 0 & 0 & | & \ast \\
0 & \blacksquare & 0 & 0 & 0 & | & \ast \\
0 & 0 & \blacksquare & 0 & 0 & | & \ast \\
0 & 0 & 0 & \blacksquare & | & \ast \\
\end{pmatrix}
\]

Now you can determine the answer.

19. Find \(h\) such that

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

is the augmented matrix of an inconsistent matrix.
Take $-3$ times the top row and add to 2 times the bottom. This yields

\[
\begin{pmatrix}
2 & h & | & 4 \\
0 & 12 - 3h & | & 2
\end{pmatrix}
\]

Now if $h = 4$ the system is inconsistent because it would have the bottom row equal to \( \begin{pmatrix} 0 & 0 & | & 2 \end{pmatrix} \).

20. 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.

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

If $h \neq 2$ then $k$ can be anything and the system represented by the augmented matrix will have a unique solution. Suppose then that $h = 2$. Then taking $-2$ times the top row and adding to the bottom row gives

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

If $k \neq 4$ there is no solution. However, if $k = 4$ you are left with the single equation, $x + 2y = 2$ and there are infinitely many solutions to this. In fact anything of the form $(2 - 2y, y)$ will work just fine.

21. Determine if the system is consistent.

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

This system is inconsistent. To see this, write the augmented matrix and do row operations. The augmented matrix is

\[
\begin{pmatrix}
1 & 2 & 1 & -1 & | & 2 \\
1 & -1 & 1 & 1 & | & 1 \\
2 & 1 & -1 & 0 & | & 1 \\
4 & 2 & 1 & 0 & | & 5
\end{pmatrix}
\]

A reduced echelon form for this matrix is

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

and the bottom row shows there is no solution.
22. Find the general solution of the system whose augmented matrix is
\[
\begin{pmatrix}
1 & 2 & 0 & | & 2 \\
1 & 3 & 4 & | & 2 \\
1 & 0 & 2 & | & 1 \\
\end{pmatrix}
\]
A reduced echelon form for this matrix is
\[
\begin{pmatrix}
1 & 0 & 0 & | & \frac{6}{25} \\
0 & 1 & 0 & | & \frac{2}{5} \\
0 & 0 & 1 & | & -\frac{3}{10} \\
\end{pmatrix}
\]
and so the solution is unique and is \( z = -1/10, y = 2/5, \) and \( x = 6/5. \)

23. Find the general solution of the system whose augmented matrix is
\[
\begin{pmatrix}
1 & 1 & 0 & | & 5 \\
1 & 0 & 3 & | & 2 \\
\end{pmatrix}
\]
A reduced echelon form for this matrix is
\[
\begin{pmatrix}
1 & 0 & 3 & | & 2 \\
0 & 1 & -3 & | & 3 \\
\end{pmatrix}
\]
and so the general solution is of the form \( y = 3 + 3z, x = 2 - 3z \) with \( z \) arbitrary.

24. Find the general solution of the system whose augmented matrix is
\[
\begin{pmatrix}
1 & 0 & 2 & 1 & 1 & | & 3 \\
0 & 1 & 0 & 4 & 2 & | & 1 \\
2 & 2 & 0 & 0 & 1 & | & 3 \\
1 & 0 & 1 & 0 & 2 & | & 2 \\
\end{pmatrix}
\]
You do the usual thing, row operations on the matrix to obtain a reduced echelon form. A reduced echelon form is
\[
\begin{pmatrix}
1 & 0 & 0 & 0 & \frac{9}{2} & | & \frac{7}{6} \\
0 & 1 & 0 & 0 & -4 & | & \frac{1}{3} \\
0 & 0 & 1 & 0 & -\frac{3}{2} & | & \frac{5}{6} \\
0 & 0 & 0 & 1 & \frac{3}{2} & | & \frac{1}{6} \\
\end{pmatrix}
\]
Therefore, the general solution is \( x_4 = 1/6 - 3/2x_5, x_3 = 5/6 + 5/2x_5, x_2 = 1/3 + 4x_5, \) and \( x_1 = 7/6 - 9/2x_5 \) with \( x_5 \) arbitrary.
Chapter 6

The \( LU \) and \( QR \) Factorization

6.0.1 Definition Of An \( LU \) Factorization

An \( LU \) factorization 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 \( LU \) factorization. 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 \( LU \) factorization.

First it should be noted not all matrices have an \( LU \) factorization and so we will emphasize the techniques for achieving it rather than formal proofs.

Example 6.0.1 Can you write \(
\begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}
\) 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 \( LU \). It has no \( LU \) factorization. This is what we mean above by saying the method lacks generality.

6.0.2 Finding An \( LU \) Factorization By Inspection

Which matrices have an \( LU \) factorization? 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 6.0.2 Find an \( LU \) factorization of \( A = \begin{pmatrix}
1 & 2 & 0 & 2 \\
1 & 3 & 2 & 1 \\
2 & 3 & 4 & 0
\end{pmatrix} \).
One way to find the LU factorization is to simply look for it directly. You need
\[
\begin{pmatrix}
1 & 2 & 0 & 2 \\
1 & 3 & 2 & 1 \\
2 & 3 & 4 & 0
\end{pmatrix}
= 
\begin{pmatrix}
a & d & h & j \\
x & 1 & 0 & 0 \\
y & z & 1
\end{pmatrix}
\begin{pmatrix}
a & 0 & 0 & 0 \\
b & 0 & e & i \\
c & 0 & 0 & f
\end{pmatrix}
.\]

Then multiplying these you get
\[
\begin{pmatrix}
 a & d & h & j \\
x a & xd + b & xh + e & xj + i \\
ya & yd + zb & yh + ze + c & yj + iz + f
\end{pmatrix}
\]
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 factorization is
\[
\begin{pmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
2 & -1 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 2 & 0 & 2 \\
0 & 1 & 2 & -1 \\
0 & 0 & 6 & -5
\end{pmatrix}
.\]

You can check whether you got it right by simply multiplying these two.

### 6.0.3 Using Multipliers To Find An LU Factorization

There is also a convenient procedure for finding an LU factorization. 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 6.0.3** Find an LU factorization 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 \(LU\) factorization is the above.

The justification for this gimmick is in my linear algebra book on the web.

Example 6.0.4 Find an \(LU\) factorization for 
\[
\begin{pmatrix}
1 & 2 & 1 & 2 & 1 \\
2 & 0 & 2 & 1 & 1 \\
2 & 3 & 1 & 3 & 2 \\
1 & 0 & 1 & 1 & 2
\end{pmatrix}
\]

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 \(LU\) factorization 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 \(LU\) factorization is of interest in solving systems of equations.
6.0.4 Solving Systems Using A LU Factorization

The reason people care about the LU factorization is it allows the quick solution of systems of equations. Here is an example.

Example 6.0.5 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 factorization and it turns out that the LU factorization can give the solution quickly. Here is how. The following is an LU factorization 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
\[
x = \begin{pmatrix}
-\frac{3}{5} + \frac{7}{5}t \\
\frac{9}{5} - \frac{11}{5}t \\
t \\
-1
\end{pmatrix}, t \in \mathbb{R}.
\]

6.1 The QR Factorization*

As pointed out above, the LU factorization is not a mathematically respectable thing because it does not always exist. There is another factorization which does always exist. Much more can be said about it than I will say here. I will only deal with real matrices and so the dot product will be the usual real dot product.
6.1. THE QR FACTORIZATION

**Definition 6.1.1** An \( n \times n \) real matrix \( Q \) is called an orthogonal matrix if

\[
QQ^T = Q^TQ = I.
\]

Thus an orthogonal matrix is one whose inverse is equal to its transpose.

First note that if a matrix is orthogonal this says

\[
\sum_j Q^T_{ij}Q_{jk} = \sum_j Q_{ji}Q_{jk} = \delta_{ik}
\]

Thus

\[
|Qx|^2 = \sum_i \left( \sum_j Q_{ij}x_j \right)^2 = \sum_i \sum_r \sum_s Q_{is}x_sQ_{ir}x_r
\]

\[
= \sum_i \sum_r \sum_s Q_{is}Q_{ir}x_sx_r = \sum_r \sum_s \sum_i Q_{is}Q_{ir}x_sx_r
\]

\[
= \sum_r \sum_s \delta_{sr}x_sx_r = \sum_r x_r^2 = |x|^2
\]

This shows that orthogonal transformations preserve distances. You can show that if you have a matrix which does preserve distances, then it must be orthogonal also.

**Example 6.1.2** One of the most important examples of an orthogonal matrix is the so-called Householder matrix. You have \( v \) a unit vector and you form the matrix,

\[
I - 2vv^T
\]

This is an orthogonal matrix which is also symmetric. To see this, you use the rules of matrix operations.

\[
(I - 2vv^T)^T = I^T - (2vv^T)^T
\]

\[
= I - 2vv^T
\]

so it is symmetric. Now to show it is orthogonal,

\[
(I - 2vv^T)(I - 2vv^T) = I - 2vv^T - 2vv^T + 4vv^Tv^Tv^T
\]

\[
= I - 4vv^T + 4vv^T = I
\]

because \( v^Tv = v \cdot v = |v|^2 = 1 \). Therefore, this is an example of an orthogonal matrix.

Consider the following problem.

**Problem 6.1.3** Given two vectors \( x, y \) such that \( |x| = |y| \neq 0 \) but \( x \neq y \) and you want an orthogonal matrix, \( Q \) such that \( Qx = y \) and \( Qy = x \). The thing which works is the Householder matrix

\[
Q \equiv I - 2 \frac{x - y}{|x - y|^2} (x - y)^T
\]

Here is why this works.

\[
Q(x - y) = (x - y) - 2 \frac{x - y}{|x - y|^2} (x - y)^T (x - y)
\]

\[
= (x - y) - 2 \frac{x - y}{|x - y|^2} |x - y|^2 = y - x
\]
\[ Q(x + y) = (x + y) - 2 \frac{x - y}{|x - y|^2} (x - y)^T (x + y) \]

\[ = (x + y) - 2 \frac{x - y}{|x - y|^2} ((x - y) \cdot (x + y)) \]

\[ = (x + y) - 2 \frac{x - y}{|x - y|^2} (|x|^2 - |y|^2) = x + y \]

Hence

\[ Qx + Qy = x + y \]
\[ Qx - Qy = y - x \]

Adding these equations, \( 2Qx = 2y \) and subtracting them yields \( 2Qy = 2x \).

A picture of the geometric significance follows.

The orthogonal matrix \( Q \) reflects across the dotted line taking \( x \) to \( y \) and \( y \) to \( x \).

**Definition 6.1.4** Let \( A \) be an \( m \times n \) matrix. Then a QR factorization of \( A \) consists of two matrices, \( Q \) orthogonal and \( R \) upper triangular or in other words equal to zero below the main diagonal such that \( A = QR \).

With the solution to this simple problem, here is how to obtain a QR factorization for any matrix \( A \). Let

\[ A = (a_1, a_2, \ldots, a_n) \]

where the \( a_i \) are the columns. If \( a_1 = 0 \), let \( Q_1 = I \). If \( a_1 \neq 0 \), let

\[ b = \begin{pmatrix} |a_1| \\ 0 \\ \vdots \\ 0 \end{pmatrix} \]

and form the Householder matrix,

\[ Q_1 = I - 2 \frac{(a_1 - b)}{|a_1 - b|^2} (a_1 - b)^T \]

As in the above problem \( Q_1a_1 = b \) and so

\[ Q_1A = \begin{pmatrix} |a_1| & * \\ 0 & A_2 \end{pmatrix} \]

where \( A_2 \) is a \( m-1 \times n-1 \) matrix. Now find in the same way as just done a \( n-1 \times n-1 \) matrix \( \hat{Q}_2 \) such that

\[ \hat{Q}_2A_2 = \begin{pmatrix} * & * \\ 0 & A_3 \end{pmatrix} \]
6.1. THE QR FACTORIZATION

Let
\[ Q_2 \equiv \begin{pmatrix} 1 & 0 \\ 0 & \tilde{Q}_2 \end{pmatrix} \]

Then
\[ Q_2Q_1A = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{Q}_2 \end{pmatrix} \begin{pmatrix} |a_1| & \ast \\ \ast & A_2 \end{pmatrix} = \begin{pmatrix} |a_1| & \ast & \ast \\ \ast & \ast & \ast \\ 0 & 0 & A_3 \end{pmatrix} \]

Continuing this way until the result is upper triangular, you get a sequence of orthogonal matrices \( Q_pQ_{p-1} \cdots Q_1 \) such that
\[ Q_pQ_{p-1} \cdots Q_1A = R \tag{6.1} \]

where \( R \) is upper triangular.

Now if \( Q_1 \) and \( Q_2 \) are orthogonal, then from properties of matrix multiplication,
\[ Q_1Q_2(Q_1Q_2)^T = Q_1Q_2Q_2^TQ_1^T = Q_1IQ_1^T = I \]

and similarly
\[ (Q_1Q_2)^TQ_1Q_2 = I. \]

Thus the product of orthogonal matrices is orthogonal. Also the transpose of an orthogonal matrix is orthogonal directly from the definition. Therefore, from \( 6.1 \)
\[ A = (Q_pQ_{p-1} \cdots Q_1)^T R \equiv QR. \]

This proves the following theorem.

**Theorem 6.1.5** Let \( A \) be any real \( m \times n \) matrix. Then there exists an orthogonal matrix, \( Q \) and an upper triangular matrix \( R \) such that
\[ A = QR \]

and this factorization can be accomplished in a systematic manner.
Chapter 7
Determinants

Quiz

1. Here is a matrix.
\[
\begin{pmatrix}
1 & 2 & 1 & 0 \\
1 & -1 & 1 & 1 \\
3 & 3 & 3 & 1
\end{pmatrix}
\]
Find its rank. Also identify the pivot columns.

2. What 3 \times 3 matrix will produce the row operation of taking 3 times the first row and adding it to the third row when multiplied on the left of a 3 \times 5 matrix?

3. Suppose you have an \(m \times n\) matrix and you multiply on the right by an \(n \times n\) matrix. What will be the effect on the given \(m \times n\) matrix?

7.1 Basic Techniques And Properties

7.1.1 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 2×2 matrix, this number is very easy to find.

Definition 7.1.1 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
\[\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix}.\]

Example 7.1.2 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 2×2 matrix, what about a 3×3 matrix?
Definition 7.1.3 Suppose $A$ is a $3 \times 3$ matrix. The \textit{$ij^{th}$ 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 7.1.4 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 7.1.5 Suppose $A$ is a $3 \times 3$ matrix. The \textit{$ij^{th}$ 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 7.1.6 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,

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

Similarly,

$$\text{cof} (A)_{22} = (-1)^{2+2} \det \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix} = -8.$$ 

**Definition 7.1.7** 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 7.1.8** 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”.

$$\begin{align*}
\text{cof} (A)_{11} &= 1(-1)^{1+1} \begin{vmatrix} 2 & 3 \\ 2 & 1 \end{vmatrix} + 4(-1)^{2+1} \begin{vmatrix} 2 & 1 \\ 3 & 1 \end{vmatrix} + 3(-1)^{3+1} \begin{vmatrix} 1 & 2 \\ 3 & 2 \end{vmatrix} = 0.
\end{align*}$$

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.

$$\begin{align*}
\text{cof} (A)_{21} &= 4(-1)^{2+1} \begin{vmatrix} 2 & 3 \\ 2 & 1 \end{vmatrix} + 3(-1)^{2+2} \begin{vmatrix} 1 & 3 \\ 3 & 1 \end{vmatrix} + 2(-1)^{2+3} \begin{vmatrix} 1 & 2 \\ 3 & 2 \end{vmatrix} = 0.
\end{align*}$$

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 now how to find the determinant of a $3 \times 3$ matrix. The pattern is the same.

**Definition 7.1.9** 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 $(-1)^{i+j} \times (ij^{th} \text{ minor})$. In words, you multiply $(-1)^{i+j}$ times the $ij^{th}$ minor to get the $ij^{th}$ cofactor.

**Definition 7.1.10** 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 7.1.11 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. Let’s pick the third column.

\[
\det (A) = 3 (-1)^{1+3} \begin{vmatrix} 1 & 3 & 5 \\ 3 & 4 & 2 \end{vmatrix} + 2 (-1)^{2+3} \begin{vmatrix} 1 & 2 & 4 \\ 3 & 4 & 2 \end{vmatrix} + 3 (-1)^{3+3} \begin{vmatrix} 1 & 2 & 4 \\ 3 & 4 & 2 \end{vmatrix} + \begin{vmatrix} 1 & 3 & 5 \\ 3 & 4 & 2 \end{vmatrix}.
\]

Now you know how to expand each of these \( 3 \times 3 \) 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 \( 2 \times 2 \) matrices and each of these will need 2 terms. Therefore, there will be \( 4 \times 3 \times 2 = 24 \) terms to evaluate in order to find the determinant using the method of Laplace expansion. Suppose now you have a \( 10 \times 10 \) 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 7.1.12 Let \( A = (a_{ij}) \) be an \( n \times n \) matrix and suppose the determinant of a \( (n-1) \times (n-1) \) matrix has been defined. 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 \( (-1)^{i+j} \). Thus \( (-1)^{i+j} \times (the \ ij^{th} \ minor) \) equals the \( ij^{th} \) cofactor. To make the formulas easier to remember, \( \text{cof} (A)_{ij} \) 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 \( n \times n \) matrix.

Definition 7.1.13 Let \( A \) be an \( n \times n \) matrix where \( n \geq 2 \) and suppose the determinant of an \( (n-1) \times (n-1) \) matrix has been defined. Then

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

The first formula consists of expanding the determinant along the $i^{th}$ row and the second expands the determinant along the $j^{th}$ column.

**Theorem 7.1.14** Expanding the $n \times n$ matrix along any row or column always gives the same answer so the above definition is a good definition.

### 7.1.2 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.

**Definition 7.1.15** 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.

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

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

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

**Corollary 7.1.16** 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 7.1.17** 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

$$
\begin{vmatrix}
2 & 6 & 7 \\
1 & 0 & 3 & 33.7 \\
0 & 0 & -1 \\
0(-1)^{3+1} & 2 & 3 & 77 \\
0 & 0 & -1 & 0 & 3 & 33.7
\end{vmatrix}
\begin{vmatrix}
2 & 3 & 77 \\
0 & 0 & -1 \\
0(-1)^{3+1} & 2 & 6 & 7 \\
0 & 0 & -1 & 0 & 3 & 33.7
\end{vmatrix}
$$

and the only nonzero term in the expansion is

$$
\begin{vmatrix}
2 & 6 & 7 \\
1 & 0 & 3 & 33.7 \\
0 & 0 & -1
\end{vmatrix}.
$$
Now expand this along the first column to obtain
\[ 1 \times \begin{vmatrix} 3 & 33.7 \\ 0 & -1 \end{vmatrix} + 0 (-1)^{2+1} \begin{vmatrix} 6 & 7 \\ 0 & -1 \end{vmatrix} + 0 (-1)^{3+1} \begin{vmatrix} 6 & 7 \\ 3 & 33.7 \end{vmatrix} \]
\[ = 1 \times 2 \times \begin{vmatrix} 3 & 33.7 \\ 0 & -1 \end{vmatrix} \]
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.

### 7.1.3 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 7.1.18** 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 7.1.19** 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 7.1.20** Let \(A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\), \(A_1 = \begin{pmatrix} 2 & 4 \\ 3 & 4 \end{pmatrix}\). \(\det(A) = -2\), \(\det(A_1) = -4\).

**Theorem 7.1.21** 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 7.1.22** Let \(A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\) and let \(A_1 = \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix}\). \(\det(A) = -2\), \(\det(A_1) = 2\).

**Theorem 7.1.23** 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 7.1.24** Let \(A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\) and let \(A_1 = \begin{pmatrix} 1 & 2 \\ 4 & 6 \end{pmatrix}\). Thus 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 7.1.25** In Theorems 7.1.19 - 7.1.23 you can replace the word “row” with the word “column”.

There are two other major properties of determinants which do not involve row operations.
7.1. BASIC TECHNIQUES AND PROPERTIES

Theorem 7.1.26 Let $A$ and $B$ be two $n \times n$ matrices. Then

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

Also,

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

Example 7.1.27 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 $\det(A) \det(B) = 8 \times (-5) = -40.$

7.1.4 Finding Determinants Using Row Operations

Theorems 7.1.23 - 7.1.25 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 7.1.28 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 7.1.23, it has the same determinant as \( A \). Now using other row operations, 
\[
\det (B) = \left( \frac{-1}{3} \right) \det (C) \text{ where } C = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 11 & 22 \\ 0 & -3 & -8 & -13 \\ 0 & 6 & 30 & 9 \end{pmatrix}.
\]

The second row was replaced by \((-3)\) times the third row added to the second row. By Theorem 7.1.23 this didn’t change the value of the determinant. Then the last row was multiplied by \((-3)\). By Theorem 7.1.19 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 7.1.23. Finally switch the third and second rows. This causes the determinant to be multiplied by \((-1)\). Thus 
\[
\det (C) = -\det (D) \text{ where } D = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -3 & -8 & -13 \\ 0 & 0 & 11 & 22 \\ 0 & 0 & 14 & -17 \end{pmatrix}.
\]

You could do more row operations or you could note that this can be easily expanded along the first column followed by expanding the \(3 \times 3\) matrix which results along its first column. Thus 
\[
\det (D) = 1(-3) \begin{vmatrix} 11 & 22 \\ 14 & -17 \end{vmatrix} = 1485
\]
and so \(\det (C) = -1485\) and \(\det (A) = \det (B) = \left( \frac{-1}{3} \right) (-1485) = 495\).

**Example 7.1.29** *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 7.1.23 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 7.1.25 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 7.1.23 has the same determinant as the original matrix. Let's 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.

### 7.1.5 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. Also recall the definition of the cofactor matrix given in Definition 7.1.12 on Page 162. 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 7.1.30** \(A^{-1}\) 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 7.1.31** 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 7.1.23 - 7.1.25 on Page 164, 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

$$ \frac{1}{12} \begin{pmatrix} -2 & -2 & 6 \\ 4 & -2 & 0 \\ 2 & 8 & -6 \end{pmatrix}^T = \begin{pmatrix} -\frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ -\frac{1}{6} & -\frac{1}{6} & \frac{2}{3} \\ \frac{1}{2} & 0 & -\frac{1}{2} \end{pmatrix} . $$

Does it work? You should check to see if it does. When the matrices are multiplied

$$ \begin{pmatrix} -\frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ -\frac{1}{6} & -\frac{1}{6} & \frac{2}{3} \\ \frac{1}{2} & 0 & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 3 & 0 & 1 \\ 1 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} $$

and so it is correct.

**Example 7.1.32** 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

$$ 6 \begin{pmatrix} \frac{1}{3} & -\frac{1}{2} \\ \frac{2}{3} & -\frac{1}{2} \\ \frac{1}{3} & -\frac{1}{2} \end{pmatrix} - \begin{pmatrix} -\frac{1}{6} & -\frac{1}{2} \\ -\frac{5}{6} & -\frac{1}{2} \\ -\frac{5}{6} & -\frac{1}{2} \end{pmatrix} = \begin{pmatrix} -\frac{1}{6} & \frac{1}{3} \\ -\frac{5}{6} & \frac{2}{3} \\ -\frac{5}{6} & \frac{2}{3} \end{pmatrix} . $$
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 7.1.30:** From the definition of the determinant in terms of expansion along a column, and letting $(a_{ir}) = A$, if $\det(A) \neq 0$,

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

Now consider

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

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 7.1.21. However, expanding this matrix, $B_k$ along the $k^{th}$ column yields

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

Summarizing,

$$
\sum_{i=1}^n a_{ir} \cof(A)_{ik} \det(A)^{-1} = \delta_{rk} = \begin{cases}
1 & \text{if } r = k \\
0 & \text{if } r \neq k
\end{cases}
$$

Now

$$
\sum_{i=1}^n a_{ir} \cof(A)_{ik} = \sum_{i=1}^n a_{ir} \cof(A)^T_{ki}
$$

which is the $kr^{th}$ entry of cof$(A)^T$. Therefore,

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

(7.2)
Using the other formula in Definition 7.1.13 and similar reasoning,
\[
\sum_{j=1}^{n} a_{rj} \text{cof}(A)_{kj} \det(A)^{-1} = \delta_{rk}
\]

Now
\[
\sum_{j=1}^{n} a_{rj} \text{cof}(A)_{kj} = \sum_{j=1}^{n} a_{rj} \text{cof}(A)^{T}_{jk}
\]
which is the \(rk^{th}\) entry of \(A \text{cof}(A)^{T}\). Therefore,
\[
A \frac{\text{cof}(A)^{T}}{\det(A)} = I, \quad (7.3)
\]
and it follows from 7.2 and 7.3 that \(A^{-1} = (a_{ij}^{-1})\), where
\[
a_{ij}^{-1} = \text{cof}(A)_{ji} \det(A)^{-1}.
\]
In other words,
\[
A^{-1} = \frac{\text{cof}(A)^{T}}{\det(A)}.
\]

Now suppose \(A^{-1}\) exists. Then by Theorem 7.1.26,
\[
1 = \det(I) = \det(AA^{-1}) = \det(A) \det(A^{-1})
\]
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 7.1.33** 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
\[
\frac{1}{e^t} \begin{pmatrix}
1 & 0 & 0 \\
0 & e^t \cos t & e^t \sin t \\
0 & -e^t \sin t & e^t \cos t
\end{pmatrix}^T = \begin{pmatrix}
e^{-t} & 0 & 0 \\
0 & \cos t & -\sin t \\
0 & \sin t & \cos t
\end{pmatrix}.
\]
7.1.6 Cramer’s Rule

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,

\[
x_i = \frac{1}{\det(A)} \text{det} \begin{pmatrix}
\ast & \cdots & y_1 & \cdots & \ast \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
\ast & \cdots & y_n & \cdots & \ast
\end{pmatrix},
\]

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.

**Example 7.1.34** Find \( y \) in the following system of equations using Cramer’s rule.

\[
\begin{align*}
x + 3y - z &= 8 \\
2x + 4y - z &= 2 \\
x + y + z &= 11
\end{align*}
\]

To do this, you use Cramer’s rule.

\[
y = \begin{vmatrix}
1 & 8 & -1 \\
2 & 2 & -1 \\
1 & 11 & 1 \\
1 & 3 & -1 \\
2 & 4 & -1 \\
1 & 1 & 1
\end{vmatrix} = \frac{31}{2}
\]

You see where this comes from. I replaced the \( y \) column with the right side in the determinant on the top and the determinant on the bottom is just the determinant of the matrix of coefficients. Note I did not have to find the other variables in order to find \( y \). Lets find \( x \) and \( z \) now.

\[
x = \begin{vmatrix}
8 & 3 & -1 \\
2 & 4 & -1 \\
11 & 1 & 1 \\
1 & 3 & -1 \\
2 & 4 & -1 \\
1 & 1 & 1
\end{vmatrix} = \frac{43}{2}
\]
\[ z = \begin{vmatrix} 1 & 3 & 8 \\ 2 & 4 & 2 \\ 1 & 1 & 11 \\ 1 & 3 & -1 \\ 2 & 4 & -1 \\ 1 & 1 & 1 \end{vmatrix} = 17 \]
Chapter 8

The Mathematical Theory Of Determinants

8.0.7 The Function $\text{sgn}$

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

Lemma 8.0.35 There exists a function, $\text{sgn}$, which maps each ordered list of numbers from $\{1, \cdots, n\}$ to one of the three numbers, 0, 1, or $-1$ which also has the following properties.

$$
\text{sgn}_n (1, \cdots, n) = 1 \quad (8.1)
$$

$$
\text{sgn}_n (i_1, \cdots, p, \cdots, q, \cdots, i_n) = -\text{sgn}_n (i_1, \cdots, q, \cdots, p, \cdots, i_n) \quad (8.2)
$$

In words, the second property states that if two of the numbers are switched, the value of the function is multiplied by $-1$.

Also, in the case where $n > 1$ and $\{i_1, \cdots, i_n\} = \{1, \cdots, n\}$ so that every number from $\{1, \cdots, n\}$ appears in the ordered list, $(i_1, \cdots, i_n)$,

$$
\text{sgn}_n (i_1, \cdots, i_{n-\theta}, n, i_{\theta+1}, \cdots, i_n) \equiv (-1)^{n-\theta} \text{sgn}_{n-1} (i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n) \quad (8.3)
$$

where $n = i_{\theta}$ in the ordered list, $(i_1, \cdots, i_n)$.

Proof: Define $\text{sign} (x) = 1$ if $x > 0$, $-1$ if $x < 0$ and 0 if $x = 0$. If $n = 1$, there is only one list and it is just the number 1. Thus one can define $\text{sgn}_1 (1) \equiv 1$. For the general case where $n > 1$, simply define

$$
\text{sgn}_n (i_1, \cdots, i_n) \equiv \text{sign} \left( \prod_{r < s} (i_s - i_r) \right)
$$

This delivers either $-1, 1$, or 0 by definition. What about the other claims? Suppose you switch $i_p$ with $i_q$ where $p < q$ so two numbers in the ordered list $(i_1, \cdots, i_n)$ are switched. Denote the new ordered list of numbers as $(j_1, \cdots, j_n)$. Thus $j_p = i_q$ and $j_q = i_p$ and if $r \notin \{p, q\}$, $j_r = i_r$. See the following illustration.
\[
\begin{array}{lllllllllll}
i_1 & i_2 & \cdots & i_p & \cdots & i_q & \cdots & i_n \\
1 & \frac{1}{2} & \cdots & \frac{i}{p} & \cdots & \frac{i}{q} & \cdots & \frac{i}{n} \\
\end{array}
\]

Then

\[
\text{sgn}_n (j_1, \cdots, j_n) \equiv \text{sign} \left( \prod_{r<s} (j_s - j_r) \right)
\]

\[
= \text{sign} \left( \frac{\text{both } p,q}{(i_p - i_q)} \prod_{p<j<q} (i_j - i_q) \prod_{p<j<q} (i_p - i_j) \prod_{r<s, s \notin \{p,q\}} (i_s - i_r) \right)
\]

The last product consists of the product of terms which were in the un-switched product \( \prod_{r<s} (i_s - i_r) \) so produces no change in sign, while the two products in the middle both introduce \( q - p - 1 \) minus signs. Thus their product produces no change in sign. The first factor is of opposite sign to the \( i_q - i_p \) which occured in \( \text{sgn}_n (i_1, \cdots, i_n) \). Therefore, this switch introduced a minus sign and

\[
\text{sgn}_n (j_1, \cdots, j_n) = - \text{sgn}_n (i_1, \cdots, i_n)
\]

Now consider the last claim. In computing \( \text{sgn}_n (i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n) \) there will be the product of \( n - \theta \) negative terms

\[
(i_{\theta+1} - n) \cdots (i_n - n)
\]

and the other terms in the product for computing \( \text{sgn}_n (i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n) \) are those which are required to compute \( \text{sgn}_{n-\theta} (i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n) \) multiplied by terms of the form \( (n - i_j) \) which are nonnegative. It follows that

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

It is obvious that if there are repeats in the list the function gives 0.

**Lemma 8.0.36** Every ordered list of distinct numbers from \( \{1, 2, \cdots, n\} \) can be obtained from every other such ordered list by a finite number of switches. Also, \( \text{sgn}_n \) is unique.

**Proof:** This is obvious if \( n = 1 \) or 2. Suppose then that it is true for sets of \( n - 1 \) elements. Take two ordered lists of numbers, \( P_1, P_2 \). Make one switch in both to place \( n \) at the end. Call the result \( P_1^n \) and \( P_2^n \). Then using induction, there are finitely many switches in \( P_1^n \) so that it will coincide with \( P_2^n \). Now switch the \( n \) in what results to where it was in \( P_2 \).

To see \( \text{sgn}_n \) is unique, if there exist two functions, \( f \) and \( g \) both satisfying \( \Box \) and \( \Box \), you could start with \( f(1, \cdots, n) = g(1, \cdots, n) = 1 \) and applying the same sequence of switches, eventually arrive at \( f(i_1, \cdots, i_n) = g(i_1, \cdots, i_n) \). If any numbers are repeated, then \( \Box \) gives both functions are equal to zero for that ordered list.
8.1. THE DETERMINANT

Definition 8.0.37 When you have an ordered list of distinct numbers from \( \{1, 2, \ldots, n\} \), say 
\[
(i_1, \ldots, i_n),
\]
this ordered list is called a permutation. The symbol for all such permutations is \( S_n \). The number \( \text{sgn}_n (i_1, \ldots, i_n) \) is called the sign of the permutation.

A permutation can also be considered as a function from the set 
\[
\{1, 2, \ldots, n\}
\]
as follows. Let \( f(k) = i_k \). Permutations are of fundamental importance in certain areas of math. For example, it was by considering permutations that Galois was able to give a criterion for solution of polynomial equations by radicals, but this is a different direction than what is being attempted here.

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

8.1 The Determinant

Definition 8.1.1 Let \( f \) be a 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,
\[
\sum_{(k_1, k_2)} f(k_1, k_2) = f(1, 2) + f(2, 1) + f(1, 1) + f(2, 2).
\]

8.1.1 The Definition

Definition 8.1.2 Let \( (a_{ij}) = A \) denote an \( n \times n \) matrix. The determinant of \( A \), denoted by \( \det(A) \) is defined by
\[
\det(A) = \sum_{(k_1, \ldots, k_n)} \text{sgn}(k_1, \ldots, k_n) a_{1k_1} \cdots a_{nk_n}
\]
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.

8.1.2 Permuting Rows Or Columns

Let \( A \) be an \( n \times n \) matrix, \( A = (a_{ij}) \) and let \( (r_1, \ldots, r_n) \) denote an ordered list of \( n \) numbers from \( \{1, \ldots, n\} \). Let \( A(r_1, \ldots, r_n) \) denote the matrix whose \( k^{\text{th}} \) row is the \( r_k \) row of the matrix \( A \). Thus
\[
\det(A(r_1, \ldots, r_n)) = \sum_{(k_1, \ldots, k_n)} \text{sgn}(k_1, \ldots, k_n) a_{r_1k_1} \cdots a_{r_nk_n}
\]
and
\[
A(1, \ldots, n) = A.
\]
Proposition 8.1.3 Let 
\[(r_1, \ldots, r_n)\]
be an ordered list of numbers from \(\{1, \ldots, n\}\). Then
\[\text{sgn} (r_1, \ldots, r_n) \det (A) = \sum_{(k_1, \ldots, k_n)} \text{sgn} (k_1, \ldots, k_n) a_{r_1k_1} \cdots a_{r_nk_n} \quad (8.5)\]
\[= \det (A (r_1, \ldots, r_n)). \quad (8.6)\]

Proof: Let \((1, \ldots, n) = (1, \ldots, r, \ldots, s, \ldots, n)\) so \(r < s\).
\[\det (A (1, \ldots, r, \ldots, s, \ldots, n)) = \sum_{(k_1, \ldots, k_n)} \text{sgn} (k_1, \ldots, k_r, \ldots, k_s, \ldots, k_n) a_{1k_1} \cdots a_{rk_r} \cdots a_{sk_s} \cdots a_{nk_n},\]
and renaming the variables, calling \(k_s, k_r\) and \(k_r, k_s\), this equals
\[= \sum_{(k_1, \ldots, k_n)} \text{sgn} (k_1, \ldots, k_s, \ldots, k_r, \ldots, k_n) a_{1k_1} \cdots a_{rk_r} \cdots a_{sk_s} \cdots a_{nk_n}\]
\[= \sum_{(k_1, \ldots, k_n)} -\text{sgn} \left( k_1, \ldots, \underbrace{k_r, \ldots, k_s}_{\text{These get switched}}, \ldots, k_n \right) a_{1k_1} \cdots a_{sk_s} \cdots a_{rk_r} \cdots a_{nk_n}\]
\[= -\det (A (1, \ldots, s, \ldots, r, \ldots, n)). \quad (8.7)\]
Consequently,
\[\det (A (1, \ldots, s, \ldots, r, \ldots, n)) = -\det (A (1, \ldots, r, \ldots, s, \ldots, n)) = -\det (A)\]
Now letting \(A (1, \ldots, s, \ldots, r, \ldots, n)\) play the role of \(A\), and continuing in this way, switching pairs of numbers,
\[\det (A (r_1, \ldots, r_n)) = (-1)^p \det (A)\]
where it took \(p\) switches to obtain \((r_1, \ldots, r_n)\) from \((1, \ldots, n)\). By Lemma 8.1.2, this implies
\[\det (A (r_1, \ldots, r_n)) = (-1)^p \det (A) = \text{sgn} (r_1, \ldots, r_n) \det (A)\]
and proves the proposition in the case when there are no repeated numbers in the ordered list, \((r_1, \ldots, r_n)\). However, if there is a repeat, say the \(s^{th}\) row equals the \(s^{th}\) row, then the reasoning of 8.8 shows that \(\det (A (r_1, \ldots, r_n)) = 0\) and also \(\text{sgn} (r_1, \ldots, r_n) = 0\) so the formula holds in this case also. ■

Observation 8.1.4 There are \(n!\) ordered lists of distinct numbers from \(\{1, \ldots, n\}\).

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 \(\{1, \ldots, n\}\) as stated in the observation.
8.1. THE DETERMINANT

8.1.3 A Symmetric Definition

With the above, it is possible to give a more symmetric description of the determinant from which it will follow that det \((A) = \det (A^T)\).

**Corollary 8.1.5** The following formula for \(\det (A)\) is valid.

\[
\det (A) = \frac{1}{n!} \sum_{(r_1, \ldots, r_n)} \sum_{(k_1, \ldots, k_n)} \text{sgn} (r_1, \ldots, r_n) \text{sgn} (k_1, \ldots, k_n) a_{r_1k_1} \cdots a_{r_nk_n}. \tag{8.9}
\]

And also \(\det (A^T) = \det (A)\) where \(A^T\) is the transpose of \(A\). (Recall that for \(A^T = (a_{ij}^T)\), \(a_{ij}^T = a_{ji}\).)

**Proof:** From Proposition 8.1.3 if the \(r_i\) are distinct,

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

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

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

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

\[
\]

8.1.4 The Alternating Property Of The Determinant

**Corollary 8.1.6** 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 \((xa_1 + yb_1, \ldots, xa_n + yb_n)\). Then

\[
\det (A) = x \det (A_1) + y \det (A_2)
\]

where the \(i^{th}\) row of \(A_1\) is \((a_1, \ldots, a_n)\) and the \(i^{th}\) row of \(A_2\) is \((b_1, \ldots, b_n)\), 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 8.1.3 when two rows are switched, the determinant of the resulting matrix is \((-1)\) times the determinant of the original matrix. By Corollary 8.1.5 the same holds for columns because the columns of the matrix equal the rows of the transposed matrix. Thus if \(A_1\) is the matrix obtained from \(A\) by switching two columns,

\[
\det (A) = \det (A^T) = - \det (A_1^T) = - \det (A_1).
\]

If \(A\) has two equal columns or two equal rows, then switching them results in the same matrix. Therefore, \(\det (A) = - \det (A)\) and so \(\det (A) = 0\).
It remains to verify the last assertion.

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

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

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

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

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

### 8.1.5 Linear Combinations And Determinants

Linear combinations have been discussed already. However, here is a review and some new terminology.

**Definition 8.1.7** A vector \(w\) is a linear combination of the vectors \(\{v_1, \ldots, v_r\}\) if there exists scalars, \(c_1, \ldots, c_r\) such that \(w = \sum_{k=1}^r c_k v_k\). This is the same as saying

\[w \in \text{span}(v_1, \ldots, v_r).\]

The following corollary is also of great use.

**Corollary 8.1.8** 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 \(A = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix}\) 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 8.1.6, the determinant of \(A\) is zero if and only if the determinant of the matrix \(B\), which has this special column placed in the last position, equals zero. Thus \(a_n = \sum_{k=1}^r c_k a_k\) and so

\[
\det(B) = \det\left( a_1 \quad \cdots \quad a_r \quad \cdots \quad a_{n-1} \quad \sum_{k=1}^r c_k a_k \right).
\]

By Corollary 8.1.6

\[
\det(B) = \sum_{k=1}^r c_k \det\left( a_1 \quad \cdots \quad a_r \quad \cdots \quad a_{n-1} \quad a_k \right) = 0.
\]

because there are two equal columns. The case for rows follows from the fact that \(\det(A) = \det(A^T)\). \(\blacksquare\)
8.1.6 The Determinant Of A Product

Recall the following definition of matrix multiplication.

**Definition 8.1.9** 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 8.1.10** Let $A$ and $B$ be $n \times n$ matrices. Then

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

**Proof:** Let $c_{ij}$ be the $ij^{th}$ entry of $AB$. Then by Proposition 8.1.3,

$$\det(AB) =$$

$$\sum_{(k_1, \cdots, k_n)} \sgn(k_1, \cdots, k_n) c_{1k_1} \cdot \cdots \cdot c_{nk_n}$$

$$= \sum_{(k_1, \cdots, k_n)} \sgn(k_1, \cdots, k_n) \left( \sum_{r_1} a_{1r_1} b_{r_1 k_1} \right) \cdot \cdots \cdot \left( \sum_{r_n} a_{nr_n} b_{r_n k_n} \right)$$

$$= \sum_{(r_1, \cdots, r_n)} \sum_{(k_1, \cdots, k_n)} \sgn(k_1, \cdots, k_n) b_{r_1 k_1} \cdot \cdots \cdot b_{r_n k_n} (a_{1r_1} \cdot \cdots \cdot a_{nr_n})$$

$$= \sum_{(r_1, \cdots, r_n)} \sgn(r_1, \cdots, r_n) a_{1r_1} \cdot \cdots \cdot a_{nr_n} \det(B) = \det(A) \det(B). \blacksquare$$

8.1.7 Cofactor Expansions

**Lemma 8.1.11** Suppose a matrix is of the form

$$M = \begin{pmatrix} A & * \\ 0 & a \end{pmatrix}$$ \quad (8.10)

or

$$M = \begin{pmatrix} A & 0 \\ * & a \end{pmatrix}$$ \quad (8.11)

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,

$$\det(M) \equiv \sum_{(k_1, \cdots, k_n)} \sgn_n(k_1, \cdots, k_n) m_{1k_1} \cdot \cdots \cdot m_{nk_n}$$
Let \( \theta \) denote the position of \( n \) in the ordered list, \((k_1, \ldots, k_n)\) then using Lemma \ref{lem:det},
\[
\det(M) = \sum_{(k_1, \ldots, k_n)} (-1)^{n-\theta} \operatorname{sgn}_{n-1}(k_1, \ldots, k_{\theta-1}, k_{\theta+1}, \ldots, k_n) m_{1k_1} \cdots m_{nk_n}
\]

Now suppose \ref{lem:det}. 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
\[
a \sum_{(k_1, \ldots, k_{n-1})} \operatorname{sgn}_{n-1}(k_1, \ldots, k_{n-1}) m_{1k_1} \cdots m_{(n-1)k_{n-1}} = a \det(A).
\]

To get the assertion in the situation of \ref{lem:det} use Corollary \ref{cor:det} and \ref{lem:det} to write
\[
\det(M) = \det(M^T) = \det\left(\begin{bmatrix} A^T & 0 \\ * & a \end{bmatrix}\right) = a \det(A^T) = a \det(A) \quad \blacksquare
\]

In terms of the theory of determinants, arguably the most important idea is that of Laplace expansion along a row or a column. This will follow from the above definition of a determinant.

**Definition 8.1.12** Let \( A = (a_{ij}) \) be an \( n \times n \) matrix. Then a new matrix called the cofactor matrix, \( \operatorname{cof}(A) \) is defined by \( \operatorname{cof}(A) = (c_{ij}) \) 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, \( \operatorname{cof}(A)_{ij} \) 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 8.1.13** Let \( A \) be an \( n \times n \) matrix where \( n \geq 2 \). Then
\[
\det(A) = \sum_{j=1}^{n} a_{ij} \operatorname{cof}(A)_{ij} = \sum_{i=1}^{n} a_{ij} \operatorname{cof}(A)_{ij}. \tag{8.12}
\]

The first formula consists of expanding the determinant along the \( i \)th row and the second expands the determinant along the \( j \)th column.

**Proof:** Let \( (a_{i1}, \ldots, a_{in}) \) be the \( i \)th row of \( A \). Let \( B_j \) be the matrix obtained from \( A \) by leaving every row the same except the \( i \)th row which in \( B_j \) equals
\[
(0, \ldots, 0, a_{ij}, 0, \ldots, 0).
\]

Then by Corollary \ref{cor:det},
\[
\det(A) = \sum_{j=1}^{n} \det(B_j)
\]

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 \( \operatorname{cof}(A)_{ij} \equiv (-1)^{i+j} \det(A^{ij}) \). At this point, recall that from Proposition \ref{prop:det} when two rows or two columns in a matrix \( M \), are switched, this results in multiplying
the determinant of the old matrix by \(-1\) to get the determinant of the new matrix. Therefore, by Lemma 8.1.11,

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

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

Therefore,

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

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

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

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

which is the formula for expanding \(\det(A)\) along the \(i\)th column. ■

### 8.1.8 Formula For The Inverse

Note that this gives an easy way to write a formula for the inverse of an \(n \times n\) matrix.

**Theorem 8.1.14** \(A^{-1}\) exists if and only if \(\det(A) \neq 0\). If \(\det(A) \neq 0\), then

\[
a_{ij}^{-1} = \det(A)^{-1} \det(A)_{ji}
\]

for \(\det(A)_{ij}\) the \(ij\)th cofactor of \(A\).

**Proof:** By Theorem 8.1.13 and letting \((a_{ir}) = A\), if \(\det(A) \neq 0\),

\[
\sum_{i=1}^{n} a_{ir} \det(A)_{ir} \det(A)^{-1} = \det(A) \det(A)^{-1} = 1.
\]

Now consider

\[
\sum_{i=1}^{n} a_{ir} \det(A)_{ik} \det(A)^{-1}
\]

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 8.1.6. However, expanding this matrix along the \(k\)th column yields

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

Summarizing,

\[
\sum_{i=1}^{n} a_{ir} \det(A)_{ik} \det(A)^{-1} = \delta_{rk}.
\]
Using the other formula in Theorem 8.1.13, and similar reasoning,
\[ \sum_{j=1}^{n} a_{rj} \cof(A)_{kj} \det(A)^{-1} = \delta_{rk} \]
This proves that if \( \det(A) \neq 0 \), then \( A^{-1} \) exists with \( A^{-1} = (a_{ij}^{-1}) \), where \( a_{ij}^{-1} = \cof(A)_{ji} \det(A)^{-1} \).

Now suppose \( A^{-1} \) exists. Then by Theorem 8.1.10,
\[ 1 = \det(I) = \det(AA^{-1}) = \det(A) \det(A^{-1}) \]
so \( \det(A) \neq 0 \). □

The next corollary points out that if an \( n \times n \) matrix \( A \) has a right or a left inverse, then it has an inverse.

**Corollary 8.1.15** Let \( A \) be an \( n \times n \) matrix and suppose there exists an \( n \times n \) matrix \( B \) such that \( BA = I \). Then \( A^{-1} \) exists and \( A^{-1} = B \). Also, if there exists \( C \) an \( n \times n \) matrix such that \( AC = I \), then \( A^{-1} \) exists and \( A^{-1} = C \).

**Proof**: Since \( BA = I \), Theorem 8.1.10 implies
\[ \det B \det A = 1 \]
and so \( \det A \neq 0 \). Therefore from Theorem 8.1.14, \( A^{-1} \) exists. Therefore,
\[ A^{-1} = (BA)^{-1} = B(AA^{-1}) = BI = B. \]
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 8.1.14 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 \( A \).

**8.1.9 Cramer’s Rule**

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,
\[ x_i = \sum_{j=1}^{n} a_{ij}^{-1} y_j = \sum_{j=1}^{n} \frac{1}{\det(A)} \cof(A)_{ji} y_j. \]
By the formula for the expansion of a determinant along a column,
\[ x_i = \frac{1}{\det(A)} \det \begin{pmatrix} * & \cdots & y_1 & \cdots & * \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ * & \cdots & y_n & \cdots & * \end{pmatrix}, \]
where here the \( i^{th} \) column of \( A \) is replaced with the column vector \((y_1, \ldots, y_n)^T\), and the determinant of this modified matrix is taken and divided by \( \det(A) \). This formula is known as Cramer’s rule.
8.2. THE CAYLEY HAMILTON THEOREM

8.1.10 Upper Triangular Matrices

Definition 8.1.16 A matrix \( M \), is upper triangular if \( M_{ij} = 0 \) whenever \( i > j \). Thus such a matrix equals zero below the main diagonal, the entries of the form \( M_{ii} \) as shown.

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

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

With this definition, here is a simple corollary of Theorem 8.1.13.

Corollary 8.1.17 Let \( M \) be an upper (lower) triangular matrix. Then \( \det(M) \) is obtained by taking the product of the entries on the main diagonal.

8.2 The Cayley Hamilton Theorem

Definition 8.2.1 Let \( A \) be an \( n \times n \) matrix. The characteristic polynomial is defined as

\[ q_A(t) \equiv \det(tI - A) \]

and the solutions to \( p_A(t) = 0 \) are called eigenvalues. For \( A \) a matrix and \( p(t) = t^n + a_{n-1}t^{n-1} + \cdots + a_1t + a_0 \), denote by \( p(A) \) the matrix defined by

\[ p(A) \equiv A^n + a_{n-1}A^{n-1} + \cdots + a_1A + a_0I. \]

The explanation for the last term is that \( A^0 \) is interpreted as \( I \), the identity matrix.

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\(^1\). The proof in this section is not the most general proof, but works well when the field of scalars is \( \mathbb{R} \) or \( \mathbb{C} \). The following lemma will help with its proof.

Lemma 8.2.2 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 \( A_i = 0 \). \( \blacksquare \)

With the lemma, here is a simple corollary.

\(^1\)A special case was first proved by Hamilton in 1853. The general case was announced by Cayley some time later and a proof was given by Frobenius in 1878.
Corollary 8.2.3 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 \( i \). If \( A_i = B_i \) for each \( A_i, B_i \) then one can substitute an \( n \times n \) matrix \( M \) for \( \lambda \) and the identity will continue to hold.

**Proof:** Subtract and use the result of the lemma. The last claim is obvious by matching terms.

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

**Theorem 8.2.4** Let \( A \) be an \( n \times n \) matrix and let \( q(\lambda) \equiv \det(\lambda I - A) \) be the characteristic polynomial. Then \( q(A) = 0 \).

**Proof:** Let \( C(\lambda) \) equal the transpose of the cofactor matrix of \( (\lambda I - A) \) for \( |\lambda| \) large.
(If \( |\lambda| \) is large enough, then \( \lambda \) cannot be in the finite list of eigenvalues of \( A \) and so for such \( \lambda \), \( (\lambda I - A)^{-1} \) exists.) Therefore, by Theorem 8.2.4
\[
C(\lambda) = q(\lambda) (\lambda I - A)^{-1}.
\]

Say
\[
q(\lambda) = a_0 + a_1 \lambda + \cdots + \lambda^n
\]

Note that each entry in \( C(\lambda) \) is a polynomial in \( \lambda \) having degree no more than \( n - 1 \). For example, you might have something like
\[
C(\lambda) = \begin{pmatrix}
\lambda^2 - 6\lambda + 9 & 3 - \lambda & 0 \\
2\lambda - 6 & \lambda^2 - 3\lambda & 0 \\
\lambda - 1 & \lambda - 1 & \lambda^2 - 3\lambda + 2
\end{pmatrix}
\]
\[
= \begin{pmatrix}
9 & 3 & 0 \\
-6 & 0 & 0 \\
-1 & -1 & 2
\end{pmatrix} + \lambda \begin{pmatrix}
-6 & -1 & 0 \\
2 & -3 & 0 \\
1 & 1 & -3
\end{pmatrix} + \lambda^2 \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]

Therefore, collecting the terms in the general case,
\[
C(\lambda) = C_0 + C_1 \lambda + \cdots + C_{n-1} \lambda^{n-1}
\]

for \( C_j \) some \( n \times n \) matrix. Then
\[
C(\lambda)(\lambda I - A) = (C_0 + C_1 \lambda + \cdots + C_{n-1} \lambda^{n-1})(\lambda I - A) = q(\lambda) I
\]

Then multiplying out the middle term, it follows that for all \( |\lambda| \) sufficiently large,
\[
a_0 I + a_1 I \lambda + \cdots + I \lambda^n = C_0 \lambda + C_1 \lambda^2 + \cdots + C_{n-1} \lambda^n
\]
\[
- [C_0 A + C_1 AA + \cdots + C_{n-1} A \lambda^{n-1}]
\]
\[
= -C_0 A + (C_0 - C_1 A) \lambda + (C_1 - C_2 A) \lambda^2 + \cdots + (C_{n-2} - C_{n-1} A) \lambda^{n-1} + C_{n-1} \lambda^n
\]

Then, using Corollary 8.2.3, one can replace \( \lambda \) on both sides with \( A \). Then the right side is seen to equal 0. Hence the left side, \( q(A) I \) is also equal to 0. ■

It is good to keep in mind the following example when considering the above proof of the Cayley Hamilton theorem. It was shown to me by Marc van Leeuwen. If \( p(\lambda) = q(\lambda) \) for all \( \lambda \) or for all \( \lambda \) large enough where \( p(\lambda), q(\lambda) \) are polynomials having matrix coefficients,
then it is not necessarily the case that \( p(A) = q(A) \) for \( A \) a matrix of an appropriate size. Let

\[
E_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, & E_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, & N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}
\]

Then a short computation shows that for all complex \( \lambda \),

\[
(\lambda I + E_1) (\lambda I + E_2) = (\lambda^2 + \lambda) I = (\lambda I + E_2) (\lambda I + E_1)
\]

However,

\[
(NI + E_1) (NI + E_2) \neq (NI + E_2) (NI + E_1)
\]

The reason this can take place is that \( N \) fails to commute with \( E_i \). Of course a scalar commutes with any matrix so there was no difficulty in obtaining that the matrix equation held for arbitrary \( \lambda \), but this factored equation does not continue to hold if \( \lambda \) is replaced by a matrix. In the above proof of the Cayley Hamilton theorem, this issue was avoided by considering only polynomials which are of the form \( C_0 + C_1 \lambda + \cdots \) in which the polynomial identity held because the corresponding matrix coefficients were equal. However, you can also argue that in the above proof, the \( C_i \) each commute with \( A \). Nevertheless, an earlier proof of the Cayley Hamilton theorem using this approach was misleading because this issue was not made clear.

### 8.3 Exercises With Answers

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

\[
3 \begin{vmatrix} 2 & 0 & 0 \\ 2 & 1 & 1 \\ 2 & 3 & 3 \end{vmatrix} = (3) (2) \begin{vmatrix} 1 & 1 \\ 3 & 3 \end{vmatrix} = 0.
\]
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

\[
\begin{vmatrix}
0 & 6 & 0 & 5 \\
0 & 8 & 1 & 12 \\
-1 & 2 & 0 & 3 \\
0 & -27 & 0 & -44
\end{vmatrix}
\]

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 \(8.13\) 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.

\[
\begin{vmatrix}
-1 & 2 & 0 & 3 \\
0 & 8 & 1 & 12 \\
0 & 6 & 0 & 5 \\
0 & 0 & -44 + (9/2)5
\end{vmatrix}
\]
Finally, take $-6/8$ times the second row and add to the third.

\[
\begin{vmatrix}
-1 & 2 & 0 & 3 \\
0 & 8 & 1 & 12 \\
0 & 0 & -6/8 & 5 + (-6/8)(12) \\
0 & 0 & 0 & -44 + (9/2)5
\end{vmatrix}.
\]

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.

\[
\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} a & c \\ b & d \end{pmatrix}
\]

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,

\[
y = \frac{\begin{vmatrix} 2 & 3 & 1 \\ 2 & 2 & -1 \\ 1 & 1 & 2 \\ 2 & -1 & -1 \\ 1 & 0 & 2 \end{vmatrix}}{5} = \frac{5}{13}.
\]

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.

\[
\det\begin{pmatrix}
e^t & e^{-t} \cos t & e^{-t} \sin t \\
e^t & -e^{-t} \cos t - e^{-t} \sin t & -e^{-t} \sin t + e^{-t} \cos t \\
e^t & 2e^{-t} \sin t & -2e^{-t} \cos t
\end{pmatrix} = 5e^t e^{2(-t)} \cos^2 t + 5e^t e^{2(-t)} \sin^2 t = 5e^{-t}
\]
which is never equal to zero for any value of $t$ and so there is no value of $t$ 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

$$
\frac{1}{-59} \begin{pmatrix} 5 & 4 & -24 \\ 1 & -11 & 7 \\ -16 & -1 & 6 \end{pmatrix}^T = \frac{1}{-59} \begin{pmatrix} 5 & 1 & -16 \\ 4 & -11 & -1 \\ -24 & 7 & 6 \end{pmatrix}.
$$

If you check this, it does work.
Chapter 9

Vector Spaces Subspaces Bases

Quiz

1. Here is a matrix.

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

Find its determinant.

2. Use the theory of determinants to find the inverse of the matrix,

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

3. Let \( C = F^T F \) where \( F \) is an \( n \times n \) real matrix. Show \( \det(C) \geq 0 \).

4. Show that if \( A^{-1} \) exists, then \( \det(A^{-1}) = 1 / \det(A) \).

5. Suppose you have \( A = LU \) where \( A \) is an \( n \times n \) matrix and \( U \) is upper triangular while \( L \) is lower triangular with ones down the main diagonal. In other words, this is an \( LU \) factorization of \( A \). Give an easy way to find \( \det(A) \).

It is time to consider the idea of a Vector space. An example of a vector space is \( \mathbb{F}^n \) with the usual definitions of vector addition and scalar multiplication. You should verify that all the axioms in the following definition hold if the vectors are things in \( \mathbb{F}^n \). For example, you know

\[
(1, 2, 3) + (5, -1, 2) = (5, -1, 2) + (1, 2, 3) = (6, 1, 5)
\]

This verifies the first of the following list of axioms. The same would hold if you wrote your vectors as column vectors. It also holds if your vectors are really matrices. Anyway, verify all these axioms. Now there are many other things which also satisfy these axioms and when such an example exists, it is called a vector space.

9.0.1 What Is A Vector Space?

As it says in the Bible, “You shall know a tree by its fruit”. In other words, you know what it is if you know what it does. Now in Matthew, this is used to identify false prophets. A
person who clings to the notion that a vector is a little arrow or a list of numbers is like one who can only recognize a false prophet if he dresses a certain way, but false prophets don’t all look the same and according to the Bible and the Book of Mormon, you identify them by their actions not their appearance. The same is true of vectors. Get away from the simplistic notion that a vector is a little arrow or a list of numbers. You need to identify them by how they act. The precise definition of vectors and a vector space follows.

**Definition 9.0.1** A vector space is an “Abelian group” of “vectors” satisfying the axioms of an Abelian group,

\[ v + w = w + v, \]

the commutative law of addition,

\[ (v + w) + z = v + (w + z), \]

the associative law for addition,

\[ v + 0 = v, \]

the existence of an additive identity,

\[ v + (-v) = 0, \]

the existence of an additive inverse, along with a field of “scalars”, \( \mathbb{F} \) which are allowed to multiply the vectors according to the following rules. (The Greek letters denote scalars.)

\[ \alpha (v + w) = \alpha v + \alpha w, \quad (9.1) \]

\[ (\alpha + \beta) v = \alpha v + \beta v, \quad (9.2) \]

\[ \alpha (\beta v) = \alpha \beta (v), \quad (9.3) \]

\[ 1v = v. \quad (9.4) \]

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.

The reason this is so important is that vector spaces are ubiquitous in applied math physics and engineering and they really don’t look the same but once you understand them in terms of how they act, you can say certain important things about all of them at once. As just mentioned \( \mathbb{F}^n \) with the usual conventions of addition and scalar multiplication is a vector space. However, there are many others.

**Example 9.0.2** Consider \( V \) as the set of functions defined on \([0, 1]\). In what follows \( f, g, h \) will be the vectors (functions) and \( \alpha, \beta \) will be scalars (real numbers or complex numbers). The way you add vectors is as follows. The expression \( f + g \) is a vector which comes from adding these vectors \( f, g \). It is defined by the following formula.

\[ (f + g)(x) \equiv f(x) + g(x) \]
Since I have specified what happens when \((f + g)\) acts on \(x\), I have told what \(f + g\) is. Now how do you multiply by a scalar?

\[(\alpha f)(x) \equiv \alpha (f(x))\]

Now with these definitions of vector addition and scalar multiplication, \(V\) is a vector space. You should verify each axiom and be sure you understand it. You must do this yourself but here is how you go about it. Consider the distributive law.

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

and so the two functions \(\alpha (f + g)\) and \(\alpha f + \alpha g\) do the same thing to \(x\) so they are the same function. That is

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

Now you verify the other axioms.

### 9.0.2 What Is A Subspace?

If you have a vector space \(V\) and a collection of vectors \(W\) which is a subset of \(V\), sometimes \(W\) is itself a vector space and sometimes it is not. When it is a vector space, it is called a subspace, sometimes a subspace of \(V\). First here is a little proposition which shows how conclusions can be obtained about vectors in any vector space just from the axioms.

**Proposition 9.0.3** The \(0\) in a vector space is unique. Furthermore, the additive inverse \(-v\) of \(v\) is unique and \(0v = 0\).

**Proof:** Suppose \(0'\) is another additive identity. Then since both \(0'\) and \(0\), are additive identities,

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

Why is the additive inverse unique? Suppose \(w\) and \(-v\) both work as additive inverses for \(v\). Then from associative law,

\[w = 0 + w = (-(v) + v) + w = -v + (v + w) = -v + 0 = -v\]

What about the assertion that \(0v = 0\)? This comes from the distributive law.

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

Now add \(-(0v)\) to both sides. This yields

\[0 = -(0v) + 0v = -(0v + 0v) = (-0v + 0v) + 0v = 0 + 0v = 0v\]

This proves the proposition.

**Definition 9.0.4** 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 \(x, y \in W\). (This will be shown below.) The span of a set of vectors as just described is an example of a subspace.
Proposition 9.0.5 Suppose \( W \) is a nonempty subset of \( V \) a vector space and whenever \( a, b \) are scalars and \( x, y \in W \) it follows
\[
ax + by \in W
\]
Then \( W \) is a subspace of \( V \).

Proof: I need to verify \( W \) is a vector space. All the axioms like commutitive property and associative property etc. hold because \( W \) is a subset of \( V \) and these things all hold for vectors in \( V \). It only remains to verify that \( 0 \in W \) and that if you add two things in \( W \) you get something in \( W \) and if you multiply something in \( W \) by a scalar, the result is something in \( W \).

First of all pick \( x \in W \). Then by assumption \( 0x + 0x \in W \). However, from Proposition 9.0.3 again,
\[
0x + 0x = 0 + 0 = 0
\]
Thus \( 0 \in W \). Now suppose \( x, y \) are in \( W \). Then by assumption and the axioms,
\[
1x + 1y = x + y \in W
\]
so when you add two vectors in \( W \), you get a vector in \( W \). Next let \( x \in W \) and \( a \) is a scalar. Then by Proposition 9.0.3 again
\[
ax = ax + 0 = ax + 0x \in W.
\]
This proves the proposition.

Example 9.0.6 Let \( V \) be the set of functions defined on \([0, 1]\). A subspace of \( V \) would be the collection of all functions which are also polynomials. Is this really a subspace? If \( p, q \) are two polynomials and \( a, b \) are scalars (numbers) is it true that
\[
ap + bq
\]
is also a polynomial? Of course. You should verify this.

Example 9.0.7 Let \( V = \mathbb{F}^n \) and let \( A \) be an \( m \times n \) matrix. Let
\[
W \equiv \{ x \in \mathbb{F}^n : Ax = 0 \}
\]
In words, \( W \) consists of all those vectors in \( \mathbb{F}^n \) which are sent to \( 0 \) when multiplied on the left by \( A \). This is a subspace.

The reason \( W \) is a subspace is this. Suppose \( x, y \in W \) and \( a, b \) are scalars. Is \( ax + by \in W \)? Is
\[
A (ax + by) = 0?
\]
Yes. This follows because
\[
A (ax + by) = aAx + bAy = a0 + b0 = 0.
\]
This particular example is so important it has a name. It is called \( N(A) \), the null space of \( A \). It is sometimes also called \( \text{ker} (A) \), the kernel of \( A \).
Example 9.0.8 Let $A$ be the matrix

$$A = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & 1 \\
2 & 0 & 2 & 0 & 2 \\
5 & 1 & 5 & 1 & 5
\end{pmatrix}$$

Find $N(A)$.

You want to find all the vectors $x = (x_1, x_2, x_3, x_4)^T \in \mathbb{F}^5$ which this matrix sends to $0$. Set up the augmented matrix and proceed to find the row reduced echelon form. This is

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

Therefore, you need

$$x_1 = -x_3 - x_5, \quad x_2 = -x_4$$

and so the solutions to this are vectors of the form

$$\begin{pmatrix}
-x_3 - x_5 \\
-x_4 \\
x_3 \\
x_4 \\
x_5
\end{pmatrix} : x_3, x_4, x_5 \in \mathbb{F}$$

You could also write this as

$$x_3 \begin{pmatrix}
-1 \\
0 \\
1 \\
0 \\
0
\end{pmatrix} + x_4 \begin{pmatrix}
0 \\
-1 \\
0 \\
1 \\
0
\end{pmatrix} + x_5 \begin{pmatrix}
-1 \\
0 \\
0 \\
0 \\
1
\end{pmatrix} : x_3, x_4, x_5 \in \mathbb{F}$$

Note this writes the solution as a linear combination of three variables. That is what it is called when you take scalars times vectors and add them together.

Example 9.0.9 Let $v \in \mathbb{F}^n$ and $v \neq 0$. Let

$$M = \{ x \in \mathbb{F}^n : v \cdot x = 0 \}$$

Is $M$ a subspace?

Yes it is. Suppose $x, y \in M$ and $a, b \in \mathbb{F}$. Is $ax + by \in M$? Is it true that $v \cdot (ax + by) = 0$?

This is where you use the properties of the dot product to write the left side is equal to $a v \cdot x + b v \cdot y = a0 + b0 = 0$.

Thus $M$ is a subspace.
Example 9.0.10 Suppose \( M = \{ x \in \mathbb{R}^3 : x_3 \geq 0 \} \) is \( M \) a subspace?

No, this is not a subspace. Consider \( x = (1, 2, 3) \). Then this is in \( M \). If \( M \) is a subspace, then every scalar multiple of \( x \) must also be in \( M \). Multiply by the scalar \(-1\). This gives \((-1, -2, -3)\) and \(-3 < 0\) so \( M \) is not closed with respect to scalar multiplication. Therefore, it is not a subspace.

Example 9.0.11 Suppose \( M = \{ x \in \mathbb{R}^3 : x_3 + x_2 = 1 \} \) is \( M \) a subspace?

No, this is not a subspace. (1, -1, 2) and (1, 2, -1) are both in \( M \) but when you add these you get (2, 1, 1) and this is not in \( M \) because the sum of the second and third entries equals 2 not 1. However, if you replace the condition \( x_3 + x_2 = 1 \) with the condition \( x_3 + x_2 = 0 \), then it is a subspace.

Example 9.0.12 Let \( M \) be the set of functions whose derivatives exist and equal 0. Is this a subspace of the vector space of all functions on \([0, 1]\)?

Yes. This is a subspace. It consists of all functions which equal a constant. Clearly if you add constants you get a constant and if you multiply a constant by a scalar, then you get a constant.

Example 9.0.13 Let \( M \) denote all continuous functions defined on \([0, 1]\) whose integrals equal zero. Is this a subspace of the set of all functions defined on \([0, 1]\)?

Yes. If \( g, f \in M \) and \( a, b \) are scalars,

\[
\int_0^1 a f(x) + bg(x) \, dx = a \int_0^1 f(x) \, dx + b \int_0^1 g(x) \, dx = a0 + b0 = 0
\]

Thus \( M \) is closed with respect to taking such a linear combination of vectors in \( M \) and so it is a subspace.

9.0.3 What Is A Span?

The span of some vectors is the collection of vectors obtained by multiplying the given vectors by scalars and adding together these products.

Definition 9.0.14 If \( \{v_1, \cdots, v_n\} \subseteq V \), a vector space, then

\[
\text{span} \left( v_1, \cdots, v_n \right) \equiv \left\{ \sum_{i=1}^n \alpha_i v_i : \alpha_i \in F \right\}.
\]

An expression of the form

\[
\sum_{i=1}^n \alpha_i v_i
\]

where the \( \alpha_i \) are scalars is called a linear combination of the vectors \( \{v_1, \cdots, v_n\} \). Thus the span is the collection of all linear combinations of the vectors \( \{v_1, \cdots, v_n\} \).

Note that the span of some vectors usually contains infinitely many vectors but there are only finitely many vectors used in forming the linear combinations. Thus

\[
\{v_1, \cdots, v_n\}
\]
is a finite set of vectors.

\[ \text{span} (\mathbf{v}_1, \ldots, \mathbf{v}_n) \]

has infinitely many vectors in it.

\[ \text{span} (\mathbf{v}_1, \ldots, \mathbf{v}_n) \neq \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \]

Also note that when you are looking at \( \text{span} (\mathbf{v}_1, \ldots, \mathbf{v}_n) \) for \( \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \) vectors in \( \mathbb{F}^m \), you are really just looking at the vectors of the form

\[
\begin{pmatrix}
\mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \\
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n \\
\end{pmatrix}
\]

where \( \begin{pmatrix}
\mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \\
\end{pmatrix} \) is the \( m \times n \) matrix which has for its \( i^{th} \) column the vector \( \mathbf{v}_i \).

When you multiply a matrix times a vector you are really just taking a linear combination of the columns of the matrix with scalars equal to the components of the vector.

Is the span of a set of vectors a subspace? Yes it is.

**Example 9.0.15** Let \( \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \) be some vectors in \( V \). Then \( \text{span} (\mathbf{v}_1, \ldots, \mathbf{v}_n) \) is a subspace.

The reason for this is that if \( a, b \) are two scalars and

\[
\sum_{j=1}^{n} \alpha_j \mathbf{v}_j, \quad \sum_{i=1}^{n} \beta_j \mathbf{v}_j
\]

are two vectors in \( \text{span} (\mathbf{v}_1, \ldots, \mathbf{v}_n) \), then by the axioms for vector spaces,

\[
a \sum_{j=1}^{n} \alpha_j \mathbf{v}_j + b \sum_{i=1}^{n} \beta_j \mathbf{v}_j = \sum_{j=1}^{n} a\alpha_j \mathbf{v}_j + \sum_{i=1}^{n} b\beta_j \mathbf{v}_j
\]

\[
= \sum_{j=1}^{n} a\alpha_j \mathbf{v}_j + b\beta_j \mathbf{v}_j
\]

\[
= \sum_{j=1}^{n} (a\alpha_j + b\beta_j) \mathbf{v}_j
\]

\[ \in \text{span} (\mathbf{v}_1, \ldots, \mathbf{v}_n) \]

and this shows \( \text{span} (\mathbf{v}_1, \ldots, \mathbf{v}_n) \) is a subspace.

Is every subspace a span of some vectors? This is also true but will be shown later. Thus you will ultimately be able to say that subspaces are just spans of vectors.

### 9.1 Linear Independence, Bases, Dimension

If you have some vectors \( \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \) and one of them is in the span of the others, that is it equals an linear combination of the other vectors, you could delete it from the list and the smaller collection of vectors would have the same span. This is desirable because you want to have as short a list as possible. The precise consideration of these ideas involves the notion of linear independence.
**Definition 9.1.1** If \( \{v_1, \cdots, v_n\} \subseteq V \), 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 \( \{v_1, \cdots, v_n\} \) is called a basis for \( V \) if

\[
\text{span}(v_1, \cdots, v_n) = V
\]

and \( \{v_1, \cdots, v_n\} \) is linearly independent. The set of vectors is linearly dependent if it is not linearly independent. The plural of basis is "bases".

**Example 9.1.2** Here are two sets of vectors

\[
\begin{align*}
\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\} \\
\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\} \\
\left\{ \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right\}
\end{align*}
\]

(9.5)

The first set of vectors is a basis for \( \mathbb{R}^3 \). This is because it spans \( \mathbb{R}^3 \) and it is linearly independent. What about the second set of vectors?

First why is the first set of vectors a basis? It is because

\[
\begin{pmatrix} x \\ y \\ z \end{pmatrix} = x \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}
\]

so since this vector is arbitrary, this shows every vector is in the span of these three. Thus it spans \( \mathbb{R}^3 \). Is it linearly independent? Suppose

\[
x \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
\]

Then it follows each \( x, y, z \) equals 0 because the left reduces to

\[
\begin{pmatrix} x \\ y \\ z \end{pmatrix}
\]

To determine what to think about the second set of vectors, make them the columns of a matrix and obtain the row reduced echelon form.

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

(9.6)
The row reduced echelon form of this matrix is
\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]  
(9.7)

By the great and glorious lemma on row operations Lemma 5.5.5, it follows the columns in the original matrix are linearly independent because the columns in the above row reduced echelon form are. Also this shows that the columns of the matrix in 9.6 must span \( \mathbb{R}^3 \) also. This is because you could have augmented the matrix in 9.6 with another column containing an arbitrary vector and then the row reduced echelon form would give the matrix of 9.7 augmented with another vector. To illustrate,
\[
\begin{pmatrix}
1 & 2 & 1 & x \\
1 & 1 & 0 & y \\
0 & 0 & 1 & z
\end{pmatrix}
\]
is the row reduced echelon form of the system of equations for \( a, b, c \) the constants which exhibit \((x, y, z)^T\) as a linear combination of the vectors of 9.7
\[
a \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + b \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} + c \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}
\]
Now solve this system of equations by placing the matrix in row reduced echelon form
\[
\begin{pmatrix}
1 & 0 & 0 & -x + 2y + z \\
0 & 1 & 0 & -y + x - z \\
0 & 0 & 1 & z
\end{pmatrix}
\]
Thus there exists a unique solution for \( a, b, c \). Note there are two bases for \( \mathbb{R}^3 \) in this example and they had the same number of vectors in them. It turns out this will always be the case.

The next theorem goes to the very heart of the matter. You desperately need to understand this theorem.

The theorem says that if a vector space equals the span of some vectors and you have another set of vectors which is linearly independent, then there are at least as many vectors in the spanning set as there are in the linearly independent set. In words: **Spanning sets are at least as large as linearly independent sets.**

**Theorem 9.1.3** Let \( V \) be a vector space and let \( W \) be a subspace which equals \( \text{span} (u_1, \cdots, u_m) \). Also suppose \( \{v_1, \cdots, v_r\} \) is a linearly independent set of vectors of \( W \). Then \( r \leq m \).

**Proof:** Since \( \{u_1, \cdots, u_m\} \) is a spanning set of \( W \) and each \( v_j \) is in \( W \), it follows there exist scalars \( a_{ij} \) such that
\[
v_j = \sum_{i=1}^{m} a_{ij} u_i
\]
I just said that each \( v_j \) is a linear combination of the vectors \( u_i \) and named the scalar which goes with \( u_i \) in the linear combination \( a_{ij} \). I want to show \( r \leq m \). This is done by showing that \( r \) can’t be larger than \( m \). Suppose then that \( r > m \). The matrix \( A \equiv (a_{ij}) \)
has \( m \) rows and \( r \) columns. Thus there are more columns than rows. It follows that not all
these columns can be pivot columns. That is there must be free variables in the system of
equations \( A \mathbf{x} = \mathbf{0} \) and so there exists \( \mathbf{x} \in \mathbb{F}^r \) such that \( A \mathbf{x} = \mathbf{0} \) although \( \mathbf{x} \neq \mathbf{0} \). Let
\[
\mathbf{x} \equiv (x_1, \cdots, x_r)^T.
\]
Then
\[
\sum_{j=1}^r x_j v_j = \left( \sum_{j=1}^r x_j \right) \left( \sum_{i=1}^m a_{ij} u_i \right) = \sum_{i=1}^m \left( \sum_{j=1}^r a_{ij} x_j \right) u_i
\]
\[
= \sum_{i=1}^m (A \mathbf{x})_i u_i = \sum_{i=1}^m 0 u_i = \mathbf{0}
\]
even though not each \( x_k = 0 \). This violates the linear independence of the vectors \( \{v_1, \cdots, v_r\} \)
and shows that it can’t be the case that \( r > m \). Therefore, \( r \leq m \). This proves the theorem.

**Corollary 9.1.4** If \( \{u_1, \cdots, u_m\} \) and \( \{v_1, \cdots, v_n\} \) are two bases for \( V \), then \( m = n \).

**Proof:** By Theorem 9.1.3, \( m \leq n \) and \( n \leq m \).

This corollary is very important so here is another proof of it given independent of the
above theorem.

**Theorem 9.1.5** Let \( V \) be a vector space and suppose \( \{u_1, \cdots, u_k\} \) and \( \{v_1, \cdots, v_m\} \)
are two bases for \( V \). Then \( k = m \).

**Proof:** Suppose \( k > m \). Then since the vectors, \( \{u_1, \cdots, u_k\} \) span \( V \), there exist scalars,
\( c_{ij} \) such that
\[
\sum_{i=1}^m c_{ij} v_i = u_j.
\]
Therefore,
\[
\sum_{j=1}^k d_j u_j = \mathbf{0} \text{ if and only if } \sum_{j=1}^k \sum_{i=1}^m c_{ij} d_j v_i = \mathbf{0}
\]
if and only if
\[
\sum_{i=1}^m \left( \sum_{j=1}^k c_{ij} d_j \right) v_i = \mathbf{0}
\]
Now since \( \{v_1, \cdots, v_n\} \) is independent, this happens if and only if
\[
\sum_{j=1}^k c_{ij} d_j = 0, \ i = 1, 2, \cdots, m.
\]
However, this is a system of \( m \) equations in \( k \) variables, \( d_1, \cdots, d_k \) and \( m < k \). Therefore,
there exists a solution to this system of equations in which not all the \( d_j \) are equal to zero.
Recall why this is so. The augmented matrix for the system is of the form
\[
\begin{pmatrix}
C & \mathbf{0}
\end{pmatrix}
\]
where \( C \) is a matrix which has more columns than rows. Therefore, there are free variables
and hence nonzero solutions to the system of equations. However, this contradicts the linear
independence of \( \{u_1, \cdots, u_k\} \) because, as explained above, \( \sum_{j=1}^k d_j u_j = \mathbf{0} \). Similarly it
cannot happen that \( m > k \). This proves the theorem.
Definition 9.1.6 A vector space $V$ is of dimension $n$ if it has a basis consisting of $n$ vectors. This is well defined thanks to Corollary 9.1.4. It is always assumed here that $n < \infty$ and in this case, such a vector space is said to be finite dimensional.

Here is an interesting observation.

Lemma 9.1.7 Suppose $\{u_1, \cdots, u_k\}$ is linearly independent and $w \notin \text{span} \{u_1, \cdots, u_k\}$. Then

\[ \{u_1, \cdots, u_k, w\} \]

is linearly independent.

Proof: Suppose

\[ \sum_{j=1}^{k} c_j u_j + d w = 0. \]

If $d = 0$, then each $c_j = 0$ because of linear independence of the $u_j$. Suppose then that $d \neq 0$. This can’t happen because you could then divide by it and obtain $w$ as a linear combination of the $u_j$ which was given to not take place. Therefore, all the constants equal 0 and so the set of vectors is linearly independent as claimed.

Theorem 9.1.8 If $V = \text{span} \{u_1, \cdots, u_n\}$ then some subset of $\{u_1, \cdots, u_n\}$ is a basis for $V$. Also, if $\{u_1, \cdots, u_k\} \subseteq V$ is linearly independent and the vector space is finite dimensional, then the set, $\{u_1, \cdots, u_k\}$, can be enlarged to obtain a basis of $V$.

Proof: Let

\[ S = \{E \subseteq \{u_1, \cdots, u_n\} \mid \text{span} (E) = V\}. \]

For $E \in S$, let $|E|$ denote the number of elements of $E$. Let

\[ m = \min \{|E| \mid E \in S\}. \]

Thus there exist vectors

\[ \{v_1, \cdots, v_m\} \subseteq \{u_1, \cdots, u_n\} \]

such that

\[ \text{span} \{v_1, \cdots, v_m\} = V \]

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,

\[ c_1, \cdots, c_m \]

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,

\[ V = \text{span} \{v_1, \cdots, v_{k-1}, v_{k+1}, \cdots, v_m\} \]

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 by Lemma 9.1.7. 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 \( \{u_1, \ldots, u_{n+1}\} \) 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 9.1.3. Therefore, this list is a basis and this proves the theorem.

Example 9.1.9 Here are some vectors in \( \mathbb{R}^4 \)
\[
\begin{pmatrix}
1 \\
2 \\
1 \\
0
\end{pmatrix},
\begin{pmatrix}
0 \\
1 \\
0 \\
1
\end{pmatrix},
\begin{pmatrix}
1 \\
1 \\
0 \\
2
\end{pmatrix},
\begin{pmatrix}
2 \\
1 \\
0 \\
-3
\end{pmatrix},
\begin{pmatrix}
1 \\
2 \\
3 \\
4
\end{pmatrix}
\]
I know these are not linearly independent because if I make them the columns of a matrix and do row operations, not all of them can be a pivot column because there can be at most four pivot columns but there are 5 vectors. Thus at least one is a linear combination of the others. Can I pick out a basis for the span of these vectors as described in the above theorem?

The answer is yes and it is easy. Make them the columns of a matrix and row reduce to find the row reduced echelon form which reveals all linear relationships using Lemma 5.5.5. Take the matrix which has these as columns.
\[
\begin{pmatrix}
1 & 0 & 1 & 2 & 1 \\
2 & 1 & 1 & 1 & 2 \\
1 & 0 & 0 & 0 & 3 \\
0 & 1 & 2 & -3 & 4
\end{pmatrix}
\]
(9.8)
The row reduced echelon form is
\[
\begin{pmatrix}
1 & 0 & 0 & 0 & 3 \\
0 & 1 & 0 & 0 & -11/3 \\
0 & 0 & 1 & 0 & 4/3 \\
0 & 0 & 0 & 1 & -5/3
\end{pmatrix}
\]
the first four columns are obviously a basis for the span of the columns in this last matrix. Therefore, by Lemma 5.5.5 the same is true of the first four columns in (9.8). Therefore, a basis for the span of the given vectors is
\[
\left\{ \begin{pmatrix}
1 \\
2 \\
1 \\
0
\end{pmatrix},
\begin{pmatrix}
0 \\
1 \\
0 \\
1
\end{pmatrix},
\begin{pmatrix}
1 \\
1 \\
0 \\
2
\end{pmatrix},
\begin{pmatrix}
2 \\
1 \\
0 \\
-3
\end{pmatrix} \right\}
\]
Example 9.1.10 Here are a pair of linearly independent vectors

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

Enlarge this set of vectors to find a basis for \( \mathbb{R}^4 \) which is something you can do according to Theorem 9.1.8.

Here is a way to do this. You make those two vectors the first two columns of a matrix and then you add the vectors \( e_1, e_2, e_3, e_4 \) as the next four columns. Recall \( e_k \) is the vector with all zeros except a 1 in the \( k \)th position. When this matrix is placed in row reduced echelon form it will reveal a basis for \( \mathbb{R}^4 \) which will include the first two column vectors. All you have to do is find the pivot columns.

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

The row reduced echelon form is

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

It follows a basis for \( \mathbb{R}^4 \) is

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

Remember I claimed above that any subspace is the span of some vectors. Well here is why this is so.

Theorem 9.1.11 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 span \( \{v_1\} = V \), stop. \( \{v_1\} \) is a basis for \( V \). Otherwise, there exists \( v_2 \in V \) which is not in span \( \{v_1\} \). By Lemma 9.1.7 \( \{v_1, v_2\} \) is a linearly independent set of vectors. If span \( \{v_1, v_2\} = V \), stop, \( \{v_1, v_2\} \) is a basis for \( V \). If span \( \{v_1, v_2\} \neq V \), then there exists \( v_3 \not\in \text{span} \{v_1, v_2\} \) 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 Theorem 9.1.3. This proves the theorem.
9.1.1 Linear Independence For Spaces Of Functions

What if your vector space was the collection of smooth functions defined on an interval \([a,b]\)? Is there a way to verify that a finite set of these functions is linearly independent? The answer is yes. There is a version of what I am about to show you which is important.

**Definition 9.1.12** Let \(f_1, \ldots, f_n\) be smooth functions defined on an interval \([a,b]\). The Wronskian of these functions is defined as follows.

\[
W(f_1, \ldots, f_n)(x) \equiv \begin{vmatrix}
    f_1(x) & f_2(x) & \cdots & f_n(x) \\
    f'_1(x) & f'_2(x) & \cdots & f'_n(x) \\
    \vdots & \vdots & \ddots & \vdots \\
    f^{(n-1)}_1(x) & f^{(n-1)}_2(x) & \cdots & f^{(n-1)}_n(x)
\end{vmatrix}
\]

Note that to get from one row to the next, you just differentiate everything in that row. The notation \(f^{(k)}(x)\) denotes the \(k\)th derivative.

With this definition, the following is the theorem. The interesting theorem involving the Wronskian has to do with the situation where the functions are solutions of a differential equation. Then much more can be said and it is really interesting in contrast to what I am about to present.

**Theorem 9.1.13** Let \(\{f_1, \ldots, f_n\}\) be smooth functions defined on \([a,b]\). Then they are linearly independent if there exists some point \(t \in [a,b]\) where \(W(f_1, \ldots, f_n)(t) \neq 0\).

**Proof:** Form the linear combination of these vectors (functions) and suppose it equals 0. Thus

\[
a_1 f_1 + a_2 f_2 + \cdots + a_n f_n = 0
\]

The question you must answer is whether this requires each \(a_j\) to equal zero. If they all must equal 0, then this means these vectors (functions) are independent. This is what it means to be linearly independent.

Differentiate the above equation \(n - 1\) times yielding the equations

\[
\begin{pmatrix}
    a_1 f_1 + a_2 f_2 + \cdots + a_n f_n = 0 \\
    a_1 f'_1 + a_2 f'_2 + \cdots + a_n f'_n = 0 \\
    \vdots \\
    a_1 f^{(n-1)}_1 + a_2 f^{(n-1)}_2 + \cdots + a_n f^{(n-1)}_n = 0
\end{pmatrix}
\]

Now plug in \(t\). Then the above yields

\[
\begin{pmatrix}
    f_1(t) & f_2(t) & \cdots & f_n(t) \\
    f'_1(t) & f'_2(t) & \cdots & f'_n(t) \\
    \vdots & \vdots & \ddots & \vdots \\
    f^{(n-1)}_1(t) & f^{(n-1)}_2(t) & \cdots & f^{(n-1)}_n(t)
\end{pmatrix}
\begin{pmatrix}
    a_1 \\
    a_2 \\
    \vdots \\
    a_n
\end{pmatrix}
= 
\begin{pmatrix}
    0 \\
    0 \\
    \vdots \\
    0
\end{pmatrix}
\]

Since the determinant of the matrix on the left is assumed to be nonzero, it follows this matrix has an inverse and so the only solution to the above system of equations is to have each \(a_k = 0\). This proves the theorem.
9.1.2 Row Space, Column Space, And Null Space

The following definition explains what these things are.

**Definition 9.1.14** Let $A$ be an $m \times n$ matrix. The row space is the span of the rows in $\mathbb{F}^n$. The column space is the span of the columns in $\mathbb{F}^m$. The null space, denoted by $N(A)$ is given by

$$N(A) \equiv \{ x \in \mathbb{F}^n : Ax = 0 \}$$

The rank of $A$ is the dimension of the column space. As explained below, it is the same as the number of pivot columns.

The row and column space, being spans are also subspaces. The null space is also a subspace as explained in Example 9.0.7. The problem is to describe a basis for these subspaces. I will explain how to do this now by considering some examples.

**Example 9.1.15** Let

$$A = \begin{pmatrix}
1 & 2 & 3 & 6 & 5 \\
4 & -2 & 3 & 5 & 0 \\
7 & 3 & 2 & 12 & 3 \\
2 & 5 & 9 & 16 & 2 
\end{pmatrix}$$

Find a basis for the column space.

This is a very easy problem to do with the row reduced echelon form and using Lemma 5.5.5. You obtain the row reduced echelon form of the matrix and then use the lemma to find out everything immediately. The row reduced echelon form of the above matrix is

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

You can see that every column in this row reduced echelon form is obtained as a linear combination of the first second third and fifth columns and that in addition these columns form a linearly independent set of vectors. By Lemma 5.5.5, the same is true of the columns of the original matrix. Therefore, a basis for the column space of $A$ is

$$\begin{pmatrix}
1 \\
4 \\
7 \\
2 
\end{pmatrix}, \begin{pmatrix}
2 \\
-2 \\
3 \\
5 
\end{pmatrix}, \begin{pmatrix}
3 \\
3 \\
2 \\
5 
\end{pmatrix}, \begin{pmatrix}
6 \\
5 \\
12 \\
9 
\end{pmatrix}, \begin{pmatrix}
5 \\
0 \\
3 \\
2 
\end{pmatrix}$$

**Procedure 9.1.16** To find a basis for the column space of a matrix, find the row reduced echelon form, use it to identify the pivot columns of the original matrix and these are the basis for the column space of the matrix. It will always work this way because of Lemma 5.5.5.

**Example 9.1.17** Let

$$A = \begin{pmatrix}
3 & 5 & -1 & 7 & 3 \\
4 & -2 & 3 & 5 & 0 \\
7 & 3 & 2 & 12 & 3 \\
2 & 5 & 9 & 16 & 2 
\end{pmatrix}$$
Find a basis for the row space and a basis for the column space. Also find the rank of the matrix.

You do the same thing as the above. You find the row reduced echelon form. The pivot columns are a basis for the column space. Since row operations keep you in the row space, a basis for the row space will be the nonzero rows in the row reduced echelon form. These must be linearly independent by the construction and their span will be the span of the original rows because all the row operations are reversible. The row reduced echelon form of the above matrix is

\[
\begin{pmatrix}
1 & 0 & 0 & 1 & \frac{73}{273} \\
0 & 1 & 0 & 1 & \frac{116}{273} \\
0 & 0 & 1 & 1 & -\frac{20}{273} \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

A basis for the column space of this row reduced echelon form is the first three columns obviously. Therefore, by Lemma 5.5.5, the same is true of \( A \). Then a basis for the column space of \( A \) is

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

A basis for the row space of \( A \) is

\[
\begin{pmatrix}
0 & 0 & 1 & 1 & -\frac{20}{273} \\
0 & 1 & 0 & 1 & \frac{116}{273} \\
1 & 0 & 0 & 1 & \frac{73}{273}
\end{pmatrix}
\]

The rank of the matrix is 3 because there are three vectors in a basis for the column space, the pivot columns.

**Procedure 9.1.18** To find the basis of the row space of a matrix, obtain the row reduced echelon form of the matrix and use the nonzero rows. These are a basis for the row space.

Note that for the row space you use the nonzero rows in the row reduced echelon form but for the column space, you must use the columns in the original matrix.

This is well illustrated by the above example. The columns in the row reduced echelon form all have a 0 in the last entry. This is not true of the columns of the original matrix. Thus these two subspaces cannot be the same.

Note something else interesting about this. The last column in the row reduced echelon form is

\[
\frac{73}{273} \times \text{first column} + \frac{116}{273} \times \text{second column} + \left( -\frac{20}{273} \right) \times \text{third column}.
\]

Lemma 5.5.7 says the same will be true for the columns of the original matrix. Thus

\[
\begin{pmatrix}
3 \\
4 \\
7 \\
2
\end{pmatrix} \frac{73}{273} + \begin{pmatrix}
5 \\
-2 \\
3 \\
5
\end{pmatrix} \frac{116}{273} + \left( -\frac{20}{273} \right) \begin{pmatrix}
-1 \\
3 \\
2 \\
9
\end{pmatrix} = \begin{pmatrix}
3 \\
3 \\
0 \\
2
\end{pmatrix}
\]
9.1. LINEAR INDEPENDENCE, BASES, DIMENSION

You might not have guessed this. The row reduced echelon form is a great revealer of secret linear relationships between columns.

There is a fundamental theorem which relates the rank of a matrix to the null space of a matrix.

**Theorem 9.1.19** Let $A$ be an $m \times n$ matrix. Then

$$\text{rank}(A) + \dim(N(A)) = n$$

The expression $\dim(N(A))$ is the dimension of the null space of $A$. It is sometimes called “nullity” of $A$. If $N(A) = \{0\}$ then define $\dim(N(A)) \equiv 0$.

**Rough Idea:** You are interested in $\dim(N(A))$. This equals the number of free variables in the system of equations whose augmented matrix is

$$\begin{pmatrix} A & 0 \end{pmatrix}$$

Every variable is either a free variable or a basic (leading) variable. These basic variables correspond to the pivot columns. Also the rank equals the number of these pivot columns (basic variables) and there are $n$ variables in all. Thus

$$n = \underbrace{\text{rank}(A)}_{\text{number of basic variables}} + \underbrace{\dim(N(A))}_{\text{number of free variables}}$$

A more precise proof follows which generalizes to other situations.

**Proof:** Recall what the rank of $A$ is. It is the span of the columns of $A$ or equivalently it is $A(\mathbb{F}^n)$. First suppose $A$ is one to one. Then $N(A) = \{0\}$ and so the claim would reduce to the assertion that $\text{rank}(A) = n$. Consider $\{Ae_1, \cdots, Ae_n\}$ where the $e_j$ are a basis for $\mathbb{F}^n$. This is a linearly independent set because if

$$\sum_{j=1}^{n} c_j Ae_j = 0$$

then

$$A \left( \sum_{j=1}^{n} c_j e_j \right) = 0$$

and so since $N(A) = \{0\}$,

$$\sum_{j=1}^{n} c_j e_j = 0$$

and since the vectors $e_j$ are independent, it follows each $c_j = 0$. In addition to this, $\{Ae_1, \cdots, Ae_n\}$ spans $A(\mathbb{F}^n)$ because if $x \in \mathbb{F}^n$, then

$$x = \sum_{i=1}^{n} x_i e_i$$

and so

$$Ax = \sum_{i=1}^{n} x_i Ae_i \in \text{span}(Ae_1, \cdots, Ae_n).$$
The interesting case is where $A$ is not one to one. That is $N(A) \neq \{0\}$. By Theorem 9.1.11 there exists a basis for $A(\mathbb{F}^n)$ of the form $\{Ax_1, \cdots, Ax_r\}$. First note $\{x_1, \cdots, x_r\}$ is linearly independent. This is because if

$$\sum_{i=1}^{r} c_i x_i = 0$$

then doing $A$ to both sides,

$$\sum_{i=1}^{r} c_i Ax_i = 0$$

and since the vectors $\{Ax_1, \cdots, Ax_r\}$ are linearly independent, it follows each $c_i = 0$. Now by Theorem 9.1.11 again, there exists a basis for $N(A), \{y_1, \cdots, y_s\}$.

Let $x \in \mathbb{F}^n$. Then $Ax \in A(\mathbb{F}^n)$ and so there exists scalars $c_1, \cdots, c_r$ such that

$$Ax = \sum_{i=1}^{r} c_i Ax_i$$

Subtracting the right from the left and using $A$ is linear, one gets

$$A \left( x - \sum_{i=1}^{r} c_i x_i \right) = 0.$$ 

Therefore,

$$x - \sum_{i=1}^{r} c_i x_i \in N(A)$$

and so there exist scalars $d_j$ such that

$$x - \sum_{i=1}^{r} c_i x_i = \sum_{j=1}^{s} d_j y_j$$

Consequently

$$x = \sum_{i=1}^{r} c_i x_i + \sum_{j=1}^{s} d_j y_j$$

showing that since $x$ was arbitrary $\mathbb{F}^n = \text{span}(x_1, \cdots, x_r, y_1, \cdots, y_s)$.

It only remains to verify $\{x_1, \cdots, x_r, y_1, \cdots, y_s\}$ is linearly independent. When this is done, it follows it must be a basis. Since all bases have the same number of vectors in them, it will follow $r + s = n$. So why is this set of vectors independent? Let

$$0 = \sum_{i=1}^{r} c_i x_i + \sum_{j=1}^{s} d_j y_j \quad (9.9)$$

Is it the case that all the $c_i = 0$ and all the $d_j = 0$? This is what needs to be shown in order to verify these vectors are linearly independent. Do $A$ to both sides. Since $Ay_j = 0$ for each $j$ this implies

$$0 = \sum_{i=1}^{r} c_i Ax_i.$$
and now since the set of vectors \( \{A\mathbf{x}_1, \cdots, A\mathbf{x}_r\} \) was chosen to be independent, it follows each \( c_i = 0 \). Then from (9.9)

\[
0 = \sum_{j=1}^{s} d_j y_j
\]

and now the linear independence of the vectors \( \{y_1, \cdots, y_s\} \) shows the scalars \( d_j \) are all equal to zero also. Therefore, the set of vectors \( \{\mathbf{x}_1, \cdots, \mathbf{x}_r, y_1, \cdots, y_s\} \) is linearly independent and spans \( \mathbb{F}^n \) so it is a basis for \( \mathbb{F}^n \). However, from the definition of rank, it follows the rank of \( A \) is \( r \) and so this proves the theorem.

Now consider the problem of finding a basis for the null space of a matrix.

**Example 9.1.20** Let

\[
A = \begin{pmatrix}
1 & 2 & 3 & 0 \\
0 & 2 & 2 & 3 \\
4 & 2 & 6 & 6
\end{pmatrix}
\]

**Find a basis for the null space of \( A \).**

This is really easy. You first obtain the row reduced echelon form. In this case it is

\[
B = \begin{pmatrix}
1 & 0 & 1 & 0 \\
0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\]

Thus the rank of the matrix is 3. From Theorem 9.1.19 you know the dimension of the null space is 1. So what is the null space? Recall that to get the row reduced echelon form, you are just multiplying on the left by a sequence of elementary matrices. Thus the null space of the original matrix is the same as the null space of the row reduced echelon form \( B \). All you have to do is consider what this is. Here is the augmented matrix for finding the solution to \( B\mathbf{x} = \mathbf{0} \).

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

Thus you need

\[
x_1 = -x_3, x_2 = -x_3, x_4 = 0
\]

and so the null space is of the form

\[
\begin{pmatrix}
-x_3 \\
-x_3 \\
x_3 \\
0
\end{pmatrix} = x_3 \begin{pmatrix}
-1 \\
-1 \\
1 \\
0
\end{pmatrix}
\]

so a basis for the null space is

\[
\left\{ \begin{pmatrix}
-1 \\
-1 \\
1 \\
0
\end{pmatrix} \right\}
\]
Example 9.1.21 Find a basis for the null space of the matrix

\[ A = \begin{pmatrix} 1 & 3 & -4 & -4 \\ 7 & 0 & -7 & -7 \\ 3 & 2 & -5 & -5 \end{pmatrix} \]

You are trying to find the solutions to \( Ax = 0 \). The augmented matrix is

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

The row reduced echelon form is

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

Therefore, you need

\[ x_1 = x_3 + x_4, \quad x_2 = x_3 + x_4 \]

and so the null space consists of vectors of the form

\[ \begin{pmatrix} x_3 + x_4 \\ x_3 + x_4 \\ x_3 \\ x_4 \end{pmatrix} = x_3 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} \]

a basis for the null space is therefore,

\[ \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} \]

Note the rank of the matrix was 2 and the dimension of the null space was also 2 which verifies the conclusion of Theorem 9.1.19.

9.2 Row Column And Determinant Rank*

There are various kinds of rank of a matrix which occur in applications. The following definition gives their definitions. It turns out they are all the same as the dimension of the column space (column rank) but you do see all the following descriptions of these versions of rank. The determinant rank is often used. It is sometimes impossible to figure it out so it is a very important result that it is the same as column rank which can be computed by counting the pivot columns.

Definition 9.2.1 A submatrix of a matrix \( A \) is the rectangular array of numbers obtained by deleting some rows and columns of \( A \). Let \( A \) be an \( m \times n \) matrix. The determinant rank of the matrix equals \( r \) where \( r \) is the largest number such that some \( r \times r \) submatrix of \( A \) has a non zero determinant. The row rank is defined to be the dimension of the span of the rows, called the row space. The column rank is defined to be the dimension of the span of the columns, called the column space.
Theorem 9.2.2 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.

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, $M$.

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

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 9.2.3 The determinant rank equals the row rank.

Proof: From Theorem 9.2.2, 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 9.2.4 If $A$ has determinant rank, $r$, then there exist $r$ columns of the matrix such that every other column is a linear combination of these $r$ columns. Also the column rank equals the determinant rank.

Proof: This follows from the above by considering $A^T$. The rows of $A^T$ are the columns of $A$ and the determinant rank of $A^T$ and $A$ are the same. Therefore, from Corollary 9.2.3, column rank of $A = \text{row rank of } A^T = \text{determinant rank of } A^T = \text{determinant rank of } A$.

The following theorem is of fundamental importance and ties together many of the ideas presented above.

Theorem 9.2.5 Let $A$ be an $n \times n$ matrix. Then the following are equivalent.

1. $\det(A) = 0$. 

2. $A$ has determinant rank $n$.

3. $A$ is invertible.

4. $A$ is row-equivalent to the identity matrix.

5. $A$ is column-equivalent to the identity matrix.

6. $A$ is row-reduced.

7. $A$ is in echelon form.

8. $A$ is in upper triangular form.

9. $A$ is in lower triangular form.

10. $A$ is in reduced echelon form.

11. $A$ is in reduced row echelon form.

12. $A$ is in reduced column echelon form.

13. $A$ is in reduced upper triangular form.

14. $A$ is in reduced lower triangular form.

15. $A$ is in reduced echelon form.

16. $A$ is in reduced row echelon form.

17. $A$ is in reduced column echelon form.

18. $A$ is in reduced upper triangular form.

19. $A$ is in reduced lower triangular form.

20. $A$ is in reduced echelon form.

21. $A$ is in reduced row echelon form.

22. $A$ is in reduced column echelon form.

23. $A$ is in reduced upper triangular form.

24. $A$ is in reduced lower triangular form.

25. $A$ is in reduced echelon form.

26. $A$ is in reduced row echelon form.

27. $A$ is in reduced column echelon form.

28. $A$ is in reduced upper triangular form.

29. $A$ is in reduced lower triangular form.

30. $A$ is in reduced echelon form.

31. $A$ is in reduced row echelon form.

32. $A$ is in reduced column echelon form.

33. $A$ is in reduced upper triangular form.

34. $A$ is in reduced lower triangular form.

35. $A$ is in reduced echelon form.

36. $A$ is in reduced row echelon form.

37. $A$ is in reduced column echelon form.

38. $A$ is in reduced upper triangular form.

39. $A$ is in reduced lower triangular form.

40. $A$ is in reduced echelon form.

41. $A$ is in reduced row echelon form.

42. $A$ is in reduced column echelon form.

43. $A$ is in reduced upper triangular form.

44. $A$ is in reduced lower triangular form.

45. $A$ is in reduced echelon form.

46. $A$ is in reduced row echelon form.

47. $A$ is in reduced column echelon form.

48. $A$ is in reduced upper triangular form.

49. $A$ is in reduced lower triangular form.

50. $A$ is in reduced echelon form.

51. $A$ is in reduced row echelon form.

52. $A$ is in reduced column echelon form.

53. $A$ is in reduced upper triangular form.

54. $A$ is in reduced lower triangular form.

55. $A$ is in reduced echelon form.

56. $A$ is in reduced row echelon form.

57. $A$ is in reduced column echelon form.

58. $A$ is in reduced upper triangular form.

59. $A$ is in reduced lower triangular form.

60. $A$ is in reduced echelon form.

61. $A$ is in reduced row echelon form.

62. $A$ is in reduced column echelon form.

63. $A$ is in reduced upper triangular form.

64. $A$ is in reduced lower triangular form.

65. $A$ is in reduced echelon form.

66. $A$ is in reduced row echelon form.

67. $A$ is in reduced column echelon form.

68. $A$ is in reduced upper triangular form.

69. $A$ is in reduced lower triangular form.

70. $A$ is in reduced echelon form.

71. $A$ is in reduced row echelon form.

72. $A$ is in reduced column echelon form.

73. $A$ is in reduced upper triangular form.

74. $A$ is in reduced lower triangular form.

75. $A$ is in reduced echelon form.

76. $A$ is in reduced row echelon form.

77. $A$ is in reduced column echelon form.

78. $A$ is in reduced upper triangular form.

79. $A$ is in reduced lower triangular form.

80. $A$ is in reduced echelon form.

81. $A$ is in reduced row echelon form.

82. $A$ is in reduced column echelon form.

83. $A$ is in reduced upper triangular form.

84. $A$ is in reduced lower triangular form.

85. $A$ is in reduced echelon form.

86. $A$ is in reduced row echelon form.

87. $A$ is in reduced column echelon form.

88. $A$ is in reduced upper triangular form.

89. $A$ is in reduced lower triangular form.

90. $A$ is in reduced echelon form.

91. $A$ is in reduced row echelon form.

92. $A$ is in reduced column echelon form.

93. $A$ is in reduced upper triangular form.

94. $A$ is in reduced lower triangular form.

95. $A$ is in reduced echelon form.

96. $A$ is in reduced row echelon form.

97. $A$ is in reduced column echelon form.

98. $A$ is in reduced upper triangular form.

99. $A$ is in reduced lower triangular form.

100. $A$ is in reduced echelon form.

101. $A$ is in reduced row echelon form.

102. $A$ is in reduced column echelon form.

103. $A$ is in reduced upper triangular form.

104. $A$ is in reduced lower triangular form.

105. $A$ is in reduced echelon form.

106. $A$ is in reduced row echelon form.

107. $A$ is in reduced column echelon form.

108. $A$ is in reduced upper triangular form.

109. $A$ is in reduced lower triangular form.

110. $A$ is in reduced echelon form.

111. $A$ is in reduced row echelon form.

112. $A$ is in reduced column echelon form.

113. $A$ is in reduced upper triangular form.

114. $A$ is in reduced lower triangular form.

115. $A$ is in reduced echelon form.

116. $A$ is in reduced row echelon form.

117. $A$ is in reduced column echelon form.

118. $A$ is in reduced upper triangular form.

119. $A$ is in reduced lower triangular form.

120. $A$ is in reduced echelon form.

121. $A$ is in reduced row echelon form.

122. $A$ is in reduced column echelon form.

123. $A$ is in reduced upper triangular form.

124. $A$ is in reduced lower triangular form.

125. $A$ is in reduced echelon form.

126. $A$ is in reduced row echelon form.

127. $A$ is in reduced column echelon form.

128. $A$ is in reduced upper triangular form.

129. $A$ is in reduced lower triangular form.

130. $A$ is in reduced echelon form.

131. $A$ is in reduced row echelon form.
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 9.2.2. In particular, it follows that for some $m$, the $m^{th}$ column is a linear combination of all the others. Thus letting $A = \begin{pmatrix} a_1 & \cdots & a_m & \cdots & a_n \end{pmatrix}$ where the columns are denoted by $a_i$, there exists scalars, $\alpha$, such that

$$a_m = \sum_{k \neq m} \alpha_k a_k.$$  

Now consider the column vector, $x \equiv \begin{pmatrix} \alpha_1 & \cdots & -1 & \cdots & \alpha_n \end{pmatrix}^T$. Then

$$Ax = -a_m + \sum_{k \neq m} \alpha_k a_k = 0.$$  

Since also $A0 = 0$, it follows $A$ is not one to one. Similarly, $A^T$ is not one to one by the same argument applied to $A^T$. This verifies that 1) implies 2).

Now suppose 2). Then since $A^T$ is not one to one, it follows there exists $x \neq 0$ such that

$$A^T x = 0.$$  

Taking the transpose of both sides yields

$$x^T A = 0$$  

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

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

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

**Corollary 9.2.6** Let $A$ be an $n \times n$ matrix. Then the following are equivalent.

1. $\det(A) \neq 0$.

2. $A$ and $A^T$ are one to one.

3. $A$ is onto.

**Proof:** This follows immediately from the above theorem.
Chapter 10

Linear Transformations

10.0.1 Outcomes

A. Define linear transformation. Interpret a matrix as a linear transformation.

B. Find a matrix that represents a linear transformation given by a geometric description.

C. Write the solution space of a homogeneous system as the span of a set of basis vectors. Determine the dimension of the solution space.

D. Relate the solutions of a non-homogeneous system to the solutions of a homogeneous system.

10.1 What Is A Linear Transformation?

Quiz

1. Find the rank of the matrix

\[
\begin{pmatrix}
1 & 2 & 1 & 1 \\
2 & 1 & 1 & 1 \\
2 & 7 & 3 & 3 \\
1 & 8 & 3 & 3 \\
\end{pmatrix}
\]

2. For \( A \) the above matrix, find

\[ N(A) = \ker(A). \]

That is, find its null space.

3. Here are three vectors. \[
\begin{pmatrix}
1 \\
2 \\
2 \\
\end{pmatrix}, \begin{pmatrix}
1 \\
3 \\
1 \\
\end{pmatrix}, \begin{pmatrix}
1 \\
1 \\
3 \\
\end{pmatrix} \]

Determine whether the vectors are independent. If they are independent, is the set of these three vectors a basis for \( \mathbb{R}^3 \)? Explain why or why not.

4. For \( A \) the matrix of Problem 1 find a basis for the column space of this matrix. The column space is defined as the span of the columns.
5. Here are two vectors
\[
\begin{pmatrix}
1 \\
2 \\
1 \\
0
\end{pmatrix}, \quad \begin{pmatrix}
0 \\
1 \\
1 \\
1
\end{pmatrix}
\]
They are independent. Tell why. Next extend to find a basis for \( \mathbb{R}^4 \).

6. Let \( V \) consist of polynomials of degree less than or equal to 3. Thus \( V \) consists of functions of \( x \) of the form \( a + bx + cx^2 + dx^3 \).

Explain why \( \{1, x, x^2, x^3\} \) is a basis for \( V \).

7. Let \( V \) be the space of functions defined on \( \mathbb{R} \). Explain why this vector space has no finite basis. Thus it is an infinite dimensional vector space. \textbf{Hint:} You might consider the above problem.

### 10.2 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 10.2.1** 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 into 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.

\[
\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 5 \\ -3 \end{pmatrix}, \quad \begin{pmatrix} 1 \ 2 \ 0 \\ 2 \ 1 \ 0 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix} = \begin{pmatrix} -3 \\ 0 \end{pmatrix}, \\
\begin{pmatrix} 1 \ 2 \ 0 \\ 2 \ 1 \ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 7 \\ 3 \end{pmatrix} = \begin{pmatrix} 14 \\ 7 \end{pmatrix}, \\
\begin{pmatrix} 1 \ 2 \ 0 \\ 2 \ 1 \ 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x + 2y \\ 2x + y \end{pmatrix}
\]

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 10.2.2** Let \( T : \mathbb{F}^n \rightarrow \mathbb{F}^m \) be a function. Thus for each \( \mathbf{x} \in \mathbb{F}^n \), \( T \mathbf{x} \in \mathbb{F}^m \).

Then \( T \) is a linear transformation if whenever \( \alpha, \beta \) are scalars and \( \mathbf{x}_1 \) and \( \mathbf{x}_2 \) are vectors in \( \mathbb{F}^n \),

\[
T (\alpha \mathbf{x}_1 + \beta \mathbf{x}_2) = \alpha T \mathbf{x}_1 + \beta T \mathbf{x}_2.
\]
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 on Page 110. 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.

The reason this concept is so important is there are many examples of things which are linear transformations. You might remember from calculus that the operator which consists of taking the derivative is a linear transformation. That is, if $f, g$ are functions (vectors) and $\alpha, \beta$ are numbers (scalars)

$$\frac{d}{dx}(\alpha f + \beta g) = \alpha \frac{d}{dx}f + \beta \frac{d}{dx}g$$

Another example of a linear transformation is that of rotation through an angle. For example, I may want to rotate every vector through an angle of 45 degrees. Such a rotation would achieve something like the following if applied to each vector corresponding to points on the picture which is standing upright.

More generally, denote a rotation by $T$. Why is such a transformation linear? Consider the following picture which illustrates a rotation.

To get $T(a + b)$, you can add $Ta$ and $Tb$. Here is why. If you add $Ta$ to $Tb$ you get the diagonal of the parallelogram determined by $Ta$ and $Tb$. This diagonal also results from rotating the diagonal of the parallelogram determined by $a$ and $b$. This is because
the rotation preserves all angles between the vectors as well as their lengths. In particular, it preserves the shape of this parallelogram. Thus both $Ta + Tb$ and $T(a + b)$ give the same directed line segment. Thus $T$ distributes across $+$ where $+$ refers to vector addition. Similarly, if $k$ is a number $Tk \mathbf{a} = kT \mathbf{a}$ (draw a picture) and so you can factor out scalars also. Thus rotations are an example of a linear transformation.

**Definition 10.2.3** 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 $\mathbf{x} \neq \mathbf{y}$

$$Tx \neq Ty.$$  

Equivalently, if $T(\mathbf{x}) = T(\mathbf{y})$, then $\mathbf{x} = \mathbf{y}$.

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 10.2.4** A linear transformation mapping $\mathbb{F}^n$ to $\mathbb{F}^m$ is called **onto** if whenever $\mathbf{y} \in \mathbb{F}^m$ there exists $\mathbf{x} \in \mathbb{F}^n$ such that $T(\mathbf{x}) = \mathbf{y}$.

Thus $T$ is onto if everything in $\mathbb{F}^m$ 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 $T \mathbb{F}^n$, $T(\mathbb{F}^n)$, or $\text{Im}(T)$ as the set of vectors of $\mathbb{F}^m$ which are of the form $T \mathbf{x}$ for some $\mathbf{x} \in \mathbb{F}^n$. In the case that $T$ is obtained from multiplication by an $m \times n$ matrix, $A$, it is standard to simply write $A(\mathbb{F}^n)$, $A \mathbb{F}^n$, or $\text{Im}(A)$ to denote those vectors in $\mathbb{F}^m$ which are obtained in the form $A\mathbf{x}$ for some $\mathbf{x} \in \mathbb{F}^n$.

### 10.3 Constructing The Matrix Of A Linear Transformation

It turns out that if $T$ is any linear transformation which maps $\mathbb{F}^n$ to $\mathbb{F}^m$, there is always an $m \times n$ matrix $A$ with the property that

$$Ax = Tx$$  \hspace{1cm} (10.1)

for all $\mathbf{x} \in \mathbb{F}^n$. Here is why. Suppose $T : \mathbb{F}^n \rightarrow \mathbb{F}^m$ is a linear transformation and you want to find the matrix defined by this linear transformation as described in [10.3]. Then if $\mathbf{x} \in \mathbb{F}^n$ it follows

$$\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}_i$$
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,

$$Tx = \sum_{i=1}^{n} x_i T(e_i)$$

$$= \begin{pmatrix} T(e_1) & \cdots & T(e_n) \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

$$\equiv A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

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 10.3.1** Let $T$ be a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$. Then the matrix, $A$ satisfying (10.1) is given by

$$\begin{pmatrix} T(e_1) & \cdots & T(e_n) \end{pmatrix}$$

where $Te_i$ is the $i^{th}$ column of $A$.

### 10.3.1 Rotations in $\mathbb{R}^2$

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 discussed above. Consider the problem of rotating through an angle of $\theta$.

**Example 10.3.2** 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.
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 the definition of the \( \cos, \sin \) the coordinates of \( T(e_1) \) are as shown in the picture. The coordinates of \( T(e_2) \) also follow from simple trigonometry. Thus

\[
Te_1 = \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix}, Te_2 = \begin{pmatrix} -\sin \theta \\ \cos \theta \end{pmatrix}.
\]

Therefore, from Theorem 10.3.1,

\[
A = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}
\]

For those who prefer a more algebraic approach, the definition of \((\cos(\theta), \sin(\theta))\) is as the \( x \) and \( y \) coordinates of the point \((1,0)\). Now the point of the vector from \((0,0)\) to \((0,1)\), \( e_2 \) is exactly \( \pi/2 \) further along along the unit circle. Therefore, when it is rotated through an angle of \( \theta \) the \( x \) and \( y \) coordinates are given by

\[
(x, y) = (\cos(\theta + \pi/2), \sin(\theta + \pi/2)) = (-\sin \theta, \cos \theta).
\]

**Example 10.3.3** 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 \( x \)

\[
A_{\theta+\phi}x = T_{\theta+\phi}x = T_\phi T_\theta x = A_\theta A_\phi x.
\]

Consequently, you must have

\[
A_{\theta+\phi} = \begin{pmatrix} \cos(\theta + \phi) & -\sin(\theta + \phi) \\ \sin(\theta + \phi) & \cos(\theta + \phi) \end{pmatrix} = A_\theta A_\phi
\]

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,
10.3. CONSTRUCTING THE MATRIX OF A LINEAR TRANSFORMATION

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.

**Example 10.3.4** Find the matrix of the linear transformation which is obtained by first rotating all vectors through an angle of $\pi/6$ and then reflecting through the $x$ axis.

As shown in Example 10.3.3, the matrix of the transformation which involves rotating through an angle of $\pi/6$ is

$$
\begin{pmatrix}
cos(\pi/6) & -\sin(\pi/6) \\
\sin(\pi/6) & \cos(\pi/6)
\end{pmatrix} = \begin{pmatrix}
\frac{1}{2}\sqrt{3} & -\frac{1}{2} \\
\frac{1}{2} & \frac{1}{2}\sqrt{3}
\end{pmatrix}
$$

The matrix for the transformation which reflects all vectors through the $x$ axis is

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

Therefore, the matrix of the linear transformation which first rotates through $\pi/6$ and then reflects through the $x$ axis is

$$
\begin{pmatrix}
1 & 0 \\
0 & -1
\end{pmatrix} \begin{pmatrix}
\frac{1}{2}\sqrt{3} & -\frac{1}{2} \\
\frac{1}{2} & \frac{1}{2}\sqrt{3}
\end{pmatrix} = \begin{pmatrix}
\frac{1}{2}\sqrt{3} & -\frac{1}{2} \\
-\frac{1}{2} & -\frac{1}{2}\sqrt{3}
\end{pmatrix}.
$$

10.3.2 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 $\mathbf{P}_u(v)$. This is done using the dot product as follows.

$$
\mathbf{P}_u(v) = \frac{v \cdot u}{u \cdot u} u
$$

Because of properties of the dot product, the map $v \to \mathbf{P}_u(v)$ is linear,

$$
\mathbf{P}_u(\alpha v + \beta w) = \frac{(\alpha v + \beta w) \cdot u}{u \cdot u} u = \alpha \left( \frac{v \cdot u}{u \cdot u} u \right) + \beta \left( \frac{w \cdot u}{u \cdot u} u \right) = \alpha \mathbf{P}_u(v) + \beta \mathbf{P}_u(w).
$$

**Example 10.3.5** 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 $\mathbf{P}_u(e_i)$ gives the $i^{th}$ column of the desired matrix. Therefore, it is only necessary to find

$$
\mathbf{P}_u(e_i) \equiv \frac{e_i \cdot u}{u \cdot u} u
$$
For the given vector in the example, this implies the columns of the desired matrix are
\[
\begin{pmatrix}
\frac{1}{14} & 1 \\
\frac{2}{14} & 2 \\
\frac{3}{14} & 3
\end{pmatrix},
\begin{pmatrix}
\frac{1}{14} & 1 \\
\frac{2}{14} & 2 \\
\frac{3}{14} & 3
\end{pmatrix},
\begin{pmatrix}
\frac{1}{14} & 1 \\
\frac{2}{14} & 2 \\
\frac{3}{14} & 3
\end{pmatrix}.
\]
Hence the matrix is
\[
\begin{pmatrix}
\frac{1}{14} & 1 & 2 & 3 \\
2 & 4 & 6 \\
3 & 6 & 9
\end{pmatrix}.
\]

10.3.3 Algebraically Defined Linear Transformations
I will illustrate this with some examples.

**Example 10.3.6** Let \( T: \mathbb{R}^4 \to \mathbb{R}^2 \) be given by
\[
T(\mathbf{x}) = T(x_1, x_2, x_3, x_4) \equiv \begin{pmatrix} x_1 + x_3 \\ x_2 + x_4 + x_1 \end{pmatrix}
\]
Determine whether \( T \) is linear and if it is obtain its matrix with respect to the standard basis.

Note \( T(\mathbf{x}) \) is the same as
\[
\begin{pmatrix}
1 & 0 & 1 & 0 \\
1 & 1 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{pmatrix}
\]
and so \( T \) is linear and its matrix with respect to the usual basis is given above.

**Example 10.3.7** Let \( T: \mathbb{R}^2 \to \mathbb{R}^2 \) be given by
\[
T(\mathbf{x}) = T(x_1, x_2) \equiv \begin{pmatrix} x_1 + x_2 \\ x_2 + 1 \end{pmatrix}
\]
Determine whether \( T \) is linear and if it is obtain its matrix with respect to the standard basis.

In this case \( T \) is not linear. To see this, note that
\[
T(\mathbf{0}) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}
\]
Thus \( T \) cannot be linear because if it were, it would have to send the zero vector \( \mathbf{0} \) to \( \mathbf{0} \). Here is why. For \( T \) linear,
\[
T(\mathbf{0}) = T(\mathbf{0} + \mathbf{0}) = T(\mathbf{0}) + T(\mathbf{0})
\]
Then adding \(-T(\mathbf{0})\) to both sides you find \( T(\mathbf{0}) = \mathbf{0} \).
This shows an easy way to spot functions which are not linear. If it fails to send \(0\) to \(0\), then it can’t be linear. However, it may send \(0\) to \(0\) and still not be linear. There is no substitute for the definition of what it means for a function to be linear,

\[
T(ax + by) = aT(x) + bT(y)
\]

**Example 10.3.8** Let \(T: \mathbb{R}^2 \to \mathbb{R}^2\) be given by

\[
T(x) = T(x_1, x_2) \equiv \begin{pmatrix} x_1^2 \\ x_2 + x_1 \end{pmatrix}
\]

Determine whether \(T\) is linear and if it is, obtain its matrix with respect to the standard basis.

This one is not linear because if it were, you would need to have

\[
T(a, b) + T(c, d) = T(a + c, b + d) = \begin{pmatrix} (a + c)^2 \\ a + c + b + d \end{pmatrix}
\]

But this does not take place because

\[
T(a, b) + T(c, d) = \begin{pmatrix} a^2 \\ a + b \end{pmatrix} + \begin{pmatrix} c^2 \\ c + d \end{pmatrix} = \begin{pmatrix} a^2 + c^2 \\ a + b + c + d \end{pmatrix}
\]

and it is not normally the case that \((a + c)^2\) is equal to \(a^2 + c^2\). At least this is not so for the usual field of scalars like \(\mathbb{R}\). Just let \(a = 1, c = 3\) for example. Thus this is not a linear function. You can suspect this is the case if you observe that one of the variables in the formula is raised to a power.

What do algebraically defined linear transformations look like? We have shown above, they are all of the form

\[
Tx = Ax
\]

where \(A\) is a suitable matrix. Thus if it is of this form, then it is is linear and if it is not realizable in this form, it is not linear. In general, then linear transformations given algebraically are all of the form

\[
Tx = \begin{pmatrix} \sum_j a_{1j}x_j \\ \sum_j a_{2j}x_j \\ \vdots \\ \sum_j a_{mj}x_j \end{pmatrix} = Ax
\]

where the \(ij^{th}\) entry of \(A\) is \(a_{ij}\). Note there are no constants added in to any component and there are no variables raised to a power or anything like \(\sin(x_j)\) or some other “nonlinear” function of a variable or variables. You also don’t have any collection of variables multiplied together occurring anywhere.

### 10.3.4 Matrices Which Are One To One Or Onto

**Lemma 10.3.9** Let \(A\) be an \(m \times n\) matrix. Then \(A(\mathbb{R}^n) = \text{span}(a_1, \ldots, a_n)\) where \(a_1, \ldots, a_n\) denote the columns of \(A\). In fact, for \(x = (x_1, \ldots, x_n)^T\),

\[
Ax = \sum_{k=1}^n x_k a_k.
\]
Proof: This follows from the definition of matrix multiplication in Definition \[\text{5.1.9}\] on Page \[\text{102}\].

The following is a theorem of major significance. First here is an interesting observation.

Observation 10.3.10 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 $A(x - y) = 0$ and so from the condition, $x - y = 0$ so that $x = y$. Thus $A$ is one to one.

Theorem 10.3.11 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, \{ $Ae_1, \cdots, Ae_n$ \} 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, \{ $Ae_1, \cdots, Ae_n$ \} must be a basis for $\mathbb{F}^n$. This is because it was just shown to be linearly independent and if it did not span $\mathbb{F}^n$ then there would exist $y \notin \text{span} \ (Ae_1, \cdots, Ae_n)$. By Lemma \[\ref{lem:9.1.3}\], it follows

\{ $Ae_1, \cdots, Ae_n, y$ \}

would be linearly independent even though there is a spanning set with only $n$ vectors. This contradicts Theorem \[\ref{thm:9.1.3}\]. Thus \{ $Ae_1, \cdots, Ae_n$ \} must span $\mathbb{F}^n$.

Therefore, for $y \in \mathbb{F}^n$ there exist constants, $c_k$ 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. This says the span of the columns of $A$ equals $\mathbb{F}^n$. If one of the columns is a linear combination of the others, it could be deleted and you would obtain a spanning set of vectors for $\mathbb{F}^n$ consisting of only $n - 1$ vectors. This contradicts Theorem \[\ref{thm:10.3.11}\]. Hence the columns must be an independent set of vectors.

If $Ax = 0$, then letting $x = (x_1, \cdots, x_n)^T$, it follows

$$
\sum_{i=1}^{n} x_i a_i = 0
$$
and so each $x_i = 0$. If $Ax = Ay$, then $A(x - y) = 0$ and so $x = y$. This shows $A$ is one to one.

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,

$$A(AB) - A = A(BA - I) = 0.$$

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

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & -1 \end{pmatrix}$$

(10.2)

Then

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

but

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

### 10.3.5 Similarity

Two $n \times n$ matrices $A, B$ are said to be similar, written as $A \sim B$ if there exists a matrix $P$ such that

$$A = P^{-1}BP$$

Similarity has the following properties. $A \sim A$. If $A \sim B$ then $B \sim A$. If $A \sim B$ and $B \sim C$ then $A \sim C$. These three properties are called the reflexive property, the symmetric property and the transitive property respectively. The reflexive property is obvious. Let $I$. Consider the symmetric property. Say $A \sim B$ so there exists $P$ such that $A = P^{-1}BP$. Then $PAP^{-1} = B$ so $B \sim A$. Consider the transitive property next. Since $A \sim B$, there exists a matrix $P$ such that $A = P^{-1}BP$. Since $B \sim C$ there exists a matrix $Q$ such that $B = Q^{-1}CQ$. Therefore, $A = P^{-1}BP = P^{-1}Q^{-1}CQP = (QP)^{-1}C(QP)$ and so $A \sim C$.

Be sure you can prove these properties. The reason for this is that when two matrices are similar, they end up representing the same linear transformation with respect to two possibly different bases.

### 10.3.6 The General Solution Of A Linear System

Recall the following definition which was discussed above.
Definition 10.3.12 \( T \) 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 10.3.13 Let \( T \) be a linear transformation. Define

\[
\ker(T) \equiv \{ x : Tx = 0 \}.
\]

In words, \( \ker(T) \) is called the **kernel** of \( T \). As just described, \( \ker(T) \) consists of the set of all vectors which \( T \) sends to \( 0 \). This is also called the **null space** of \( T \). It is also called the **solution space** of the equation \( Tx = 0 \).

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 10.3.14 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 \( \ker\left( \frac{df}{dx} \right) = \) 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 10.3.15 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 \( y - x_p = y + (-1)x_p \). Then \( T(y - x_p) = Ty - Tx_p = b - b = 0 \). Let \( x = y - x_p \). 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 10.3.16** Let

\[
A = \begin{pmatrix}
1 & 2 & 3 & 0 \\
2 & 1 & 1 & 2 \\
4 & 5 & 7 & 2 \\
\end{pmatrix}
\]

Find \( \ker(A) \). Equivalently, find the solution space to the system of equations \( Ax = 0 \).

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 the row reduced echelon form,

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

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,

\[
\begin{pmatrix}
\frac{1}{3}z - \frac{4}{3}w \\
\frac{2}{3}w - \frac{5}{3}z \\
z \\
w \\
\end{pmatrix}
= z \begin{pmatrix}
\frac{1}{3} \\
-\frac{5}{3} \\
1 \\
0 \\
\end{pmatrix} + w \begin{pmatrix}
-\frac{4}{3} \\
\frac{2}{3} \\
0 \\
1 \\
\end{pmatrix}
\]
Example 10.3.17 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 10.3.16. Therefore, from Theorem 10.3.15, you will obtain all solutions to the above linear system in the form
\[
z \begin{pmatrix}
\frac{1}{3} \\
-\frac{5}{3} \\
1 \\
0
\end{pmatrix}
+ w \begin{pmatrix}
-\frac{4}{3} \\
\frac{2}{3} \\
0 \\
1
\end{pmatrix}
+ \begin{pmatrix}
1 \\
1 \\
2 \\
1
\end{pmatrix}
\]
because \( \begin{pmatrix} x \\
y \\
z \\
w \end{pmatrix} = \begin{pmatrix} 1 \\
1 \\
2 \\
1 \end{pmatrix} \) is a particular solution to the given system of equations.

10.4 Exercises With Answers

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
\[
\cos \left(\frac{5\pi}{12}\right) = -\frac{1}{4} \sqrt{2} + \frac{1}{4} \sqrt{3} \sqrt{2} \approx 0.258 \ 819 \ 05
\]
\[
\sin \left(\frac{5\pi}{12}\right) = \frac{1}{4} \sqrt{3} \sqrt{2} + \frac{1}{4} \sqrt{2} \approx 0.965 \ 925 \ 83
\]

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
\[
\begin{pmatrix}
-\frac{1}{2} \\
\frac{\sqrt{3}}{2}
\end{pmatrix}
\]
and then this becomes
\[
\begin{pmatrix}
-\frac{1}{2} \\
-\frac{\sqrt{3}}{2}
\end{pmatrix}.
\]
This is the first column of the desired matrix. Next \(e_2\) first is rotated through the given angle to give
\[
\begin{pmatrix}
-\frac{\sqrt{3}}{2} \\
-\frac{1}{2}
\end{pmatrix}
\]
and then it is reflected across the \(x\) axis to give
\[
\begin{pmatrix}
-\frac{\sqrt{3}}{2} \\
\frac{1}{2}
\end{pmatrix}
\]
and this gives the second column of the desired matrix. Thus the matrix is
\[
\begin{pmatrix}
-\frac{1}{2} & -\frac{\sqrt{3}}{2} \\
-\frac{\sqrt{3}}{2} & \frac{1}{2}
\end{pmatrix}.
\]

3. Find the matrix for \(P_u(v)\) where \(u = (1, -2, 3)^T\).

Recall
\[
P_u(v) = \frac{v \cdot u}{||u||^2} u
\]

Therefore,
\[
P_u(e_1) = \frac{1}{14} \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \quad P_u(e_2) = \frac{1}{14} \begin{pmatrix} 1 \\ -2 \end{pmatrix},
\]
\[
P_u(e_2) = \frac{3}{14} \begin{pmatrix} 1 \\ -2 \end{pmatrix}.
\]

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 $T_u(v) \equiv v - P_u(v)$ is also a linear transformation.

$$T_u(\alpha v + \beta w) = \alpha v + \beta w - P_u(\alpha v + \beta w)$$

which from [3] equals

$$\alpha(v - P_u(v)) + \beta(w - P_u(w)) = \alpha T_u v + \beta T_u w.$$

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

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

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

This might not be the first thing you would think of.
Chapter 11

Eigenvalues And Eigenvectors Of A Matrix

Quiz

1. A linear transformation involves first rotating the vectors in \( \mathbb{R}^2 \) counterclockwise through an angle of 30 degrees and then reflecting across the \( x \) axis. Find the matrix of this linear transformation.

2. A linear transformation involves projecting all vectors on to the span of the vector \( (1, 1, 1) \). Find the matrix of this linear transformation.

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.

11.1 Definition Of Eigenvectors And Eigenvalues

In this section, \( \mathbb{F} = \mathbb{C} \).

To illustrate the idea behind what will be discussed, consider the following example.

Example 11.1.1 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?
CHAPTER 11. EIGENVALUES AND EIGENVECTORS OF A MATRIX

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
\[
\begin{pmatrix}
0 & 5 & -10 \\
0 & 22 & 16 \\
0 & -9 & -2
\end{pmatrix}
\begin{pmatrix}
1 \\
0 \\
0
\end{pmatrix}
= 
\begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}
= 0
\begin{pmatrix}
1 \\
0 \\
0
\end{pmatrix}.
\]

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 11.1.2** Let $M$ be an $n \times n$ matrix and let $x \in \mathbb{C}^n$ be a nonzero vector for which

\[ Mx = \lambda x \quad (11.1) \]

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 scalar multiple of $x$. Since the zero vector, $\mathbf{0}$ has no direction this would make no sense for the zero vector. As noted above, $\mathbf{0}$ is never allowed to be an eigenvector. How can eigenvectors and eigenvalues be identified?

There is an important characterization of when a matrix is invertible in terms of determinants. This is proved completely in the section on the theory of determinants where a formula is given for the inverse in terms of the determinant and cofactors.

**Theorem 11.1.3** Let $M$ be an $n \times n$ matrix and let $T_M$ denote the linear transformation determined by $M$. Thus $T_Mx = Mx$. Then the following are equivalent.

1. $T_M$ is one to one.

2. $T_M$ is onto.

3. $\det(M) \neq 0$. 
Suppose $x$ satisfies (11.1). 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,

$$x = \left( (M - \lambda I)^{-1} (M - \lambda I) \right) x = (M - \lambda I)^{-1} ((M - \lambda I) x) = (M - \lambda I)^{-1} 0 = 0,$$

and this would require $x = 0$, contrary to the requirement that $x \neq 0$. By Theorem 11.1.3, 

$$\det (M - \lambda I) = 0.$$ (11.2)

(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, (11.2) has a solution, $\lambda \in \mathbb{C}$ by the fundamental theorem of algebra. Is it actually an eigenvalue? The answer is yes by Theorem 11.1.3. Since $\lambda I - M$ has no inverse due to its determinant equaling zero, it must fail to be one to one and so there must exist a nonzero vector which it maps to zero. This proves the following corollary.

**Corollary 11.1.4** 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 $(M - \lambda I) x = 0$.

### 11.2 Finding Eigenvectors And Eigenvalues

As an example, consider the following.

**Example 11.2.1** 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

$$\det \begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = 0$$

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
\[
\begin{pmatrix}
0 & -10 & -5 & | & 0 \\
2 & 9 & 2 & | & 0 \\
-4 & -8 & 1 & | & 0
\end{pmatrix}.
\]

The row 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 (11.3) is
\[
\begin{pmatrix}
t
\end{pmatrix}
\]

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 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
11.2. FINDING EIGENVECTORS AND EIGENVALUES

vector and see if we get 5 times this vector.

\[
\begin{pmatrix}
5 & -10 & -5 \\
2 & 14 & 2 \\
-4 & -8 & 6
\end{pmatrix}
\begin{pmatrix}
5 \\
-2 \\
4
\end{pmatrix}
= 
\begin{pmatrix}
25 \\
-10 \\
20
\end{pmatrix}
= 5
\begin{pmatrix}
5 \\
-2 \\
4
\end{pmatrix}
\]

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 \( \mathbb{R}^3 \) 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, \( \mathbf{0} \) 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,

\[
\begin{pmatrix}
-5 & -10 & -5 & 0 \\
2 & 4 & 2 & 0 \\
-4 & -8 & -4 & 0
\end{pmatrix}
\]

The row 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

\[
\begin{pmatrix}
-2s - t \\
s \\
t
\end{pmatrix}
= s
\begin{pmatrix}
-2 \\
1 \\
0
\end{pmatrix}
+ t
\begin{pmatrix}
-1 \\
0 \\
1
\end{pmatrix}.
\]

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.

### 11.3 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

**Eigenvectors are never equal to zero!**

Here is another example.

**Example 11.3.1** 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.

**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 row 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 row 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 the row 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

\[
\begin{pmatrix}
1 \\
1 \\
0
\end{pmatrix}
\]

where \( t \neq 0 \).

## 11.4 Some Useful Facts

Let \( A \) be an \( n \times n \) matrix. The sum of the entries of the matrix which are on the main diagonal is called the trace of \( A \). It turns out that this always equals the sum of the eigenvalues.

Another interesting fact is that the determinant of the matrix \( A \) always equals the product of the eigenvalues. To illustrate, consider

**Example 11.4.1** Let

\[
A = \begin{pmatrix}
8 & -4 & 2 \\
1 & 0 & 3 \\
8 & -8 & 10
\end{pmatrix}
\]

Then you can find its eigenvalues are 10, 6, 2. The sum of these equals 18 which is also the trace. The determinant of this matrix equals 120 which is the product of the eigenvalues.

## 11.5 Defective And Nondefective Matrices

**Definition 11.5.1** 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 \( r_j \) is some integer no smaller than 1. Thus the eigenvalues are \( \lambda_1, \lambda_2, \cdots, \lambda_m \). The **algebraic multiplicity** of \( \lambda_j \) is defined to be \( r_j \).

**Example 11.5.2** Consider the matrix,

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

*(11.5)*

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 11.5.3 The geometric multiplicity of an eigenvalue is the dimension of the eigenspace, 
\[ \text{ker}(A - \lambda I). \]

Example 11.5.4 Find the geometric multiplicity of \( \lambda = 1 \) for the matrix in 11.5.

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 11.5.5 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 11.5.6 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 called the Jordan Canonical form which is much more difficult than everything discussed in the entire differential equations course. If you become interested in this, see a good book in linear algebra. The good ones do discuss this topic. There is such a linear algebra book on my web page.

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, 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 11.5.7** 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, $\{v_1, \ldots, v_r\}$ is linearly independent.

**Proof:** If the conclusion of this theorem is not true, then there exist non zero scalars, $c_{kj}$ such that

$$\sum_{j=1}^{m} c_{kj} v_{kj} = 0. \quad (11.6)$$

Take $m$ to be the smallest number possible for an expression of the form (11.6) to hold. Then solving for $v_{k_1}$

$$v_{k_1} = \sum_{k_j \neq k_1} d_{k_j} v_{kj} \quad (11.7)$$

where $d_{kj} = c_{kj} / c_{k_1} \neq 0$. Multiplying both sides by $M$,

$$\lambda_{k_1} v_{k_1} = \sum_{k_j \neq k_1} d_{kj} \lambda_j v_{kj}$$

which from (11.6) yields

$$\sum_{k_j \neq k_1} d_{kj} \lambda_{k_1} v_{kj} = \sum_{k_j \neq k_1} d_{kj} \lambda_j v_{kj}$$
and therefore,

\[ 0 = \sum_{k_j \neq k_1} d_{k_j} (\lambda_{k_1} - \lambda_{k_j}) v_{k_j}, \]

a sum having fewer than \( m \) terms. However, from the assumption that \( m \) is as small as possible for \( \mathbf{v} \) to hold with all the scalars, \( c_{k_j} \) non zero, it follows that for some \( j \neq 1 \),

\[ d_{k_j} (\lambda_{k_1} - \lambda_{k_j}) = 0 \]

which implies \( \lambda_{k_1} = \lambda_{k_j} \), a contradiction.

### 11.6 Diagonalization

**Definition 11.6.1** Let \( A \) be an \( n \times n \) matrix. Then \( A \) is **diagonalizable** if there exists an invertible matrix, \( S \) such that

\[ S^{-1}AS = D \]

where \( D \) is a diagonal matrix. This means \( D \) has a zero as every entry except for the main diagonal.

**Theorem 11.6.2** An \( n \times n \) matrix is diagonalizable if and only if \( \mathbb{F}^n \) has a basis of eigenvectors of \( A \). Furthermore, you can take the matrix, \( S \) described above to be given as

\[ S = \begin{pmatrix} v_1 & v_2 & \cdots & v_n \end{pmatrix} \]

where here the \( v_k \) are the eigenvectors in the basis for \( \mathbb{F}^n \). If \( A \) is diagonalizable, the eigenvalues of \( A \) are the diagonal entries of the diagonal matrix.

**Proof:** Suppose there exists a basis of eigenvectors, \( \{v_k\} \) where \( Av_k = \lambda_k v_k \). Then let \( S \) be given as above. It follows \( S^{-1} \) exists and is of the form

\[ S^{-1} = \begin{pmatrix} w_1^T \\ w_2^T \\ \vdots \\ w_n^T \end{pmatrix} \]

where \( w_k^T v_j = \delta_{kj} \). Then

\[
\begin{pmatrix} \lambda_1 & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} = \begin{pmatrix} w_1^T \\ w_2^T \\ \vdots \\ w_n^T \end{pmatrix} \begin{pmatrix} \lambda_1 v_1 & \lambda_2 v_2 & \cdots & \lambda_n v_n \end{pmatrix} = \begin{pmatrix} w_1^T \\ w_2^T \\ \vdots \\ w_n^T \end{pmatrix} \begin{pmatrix} Av_1 & Av_2 & \cdots & Av_n \end{pmatrix} = S^{-1}AS
\]
Next suppose $A$ is diagonalizable so $S^{-1}AS = D$. Let $S = \begin{pmatrix} v_1 & v_2 & \cdots & v_n \end{pmatrix}$ where the columns are the $v_k$ and

$$D = \begin{pmatrix} \lambda_1 & 0 \\ & \ddots \\ 0 & \lambda_n \end{pmatrix}$$

Then

$$AS = SD = \begin{pmatrix} v_1 & v_2 & \cdots & v_n \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ & \ddots \\ 0 & \lambda_n \end{pmatrix}$$

and so

$$\begin{pmatrix} Av_1 & Av_2 & \cdots & Av_n \end{pmatrix} = \begin{pmatrix} \lambda_1v_1 & \lambda_2v_2 & \cdots & \lambda_nv_n \end{pmatrix}$$

showing the $v_i$ are eigenvectors of $A$ and the $\lambda_k$ are eigenvectors. Now the $v_k$ form a basis for $\mathbb{F}^n$ because the matrix, $S$ having these vectors as columns is given to be invertible. This proves the theorem.

**Definition 11.6.3** Let $A, B$ be two diagonal matrices. Then $A$ is said to be similar to $B$ if there exists an invertible matrix, $S$ such that $B = S^{-1}AS$.

**Example 11.6.4** Let $A = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 4 & -1 \\ -2 & -4 & 4 \end{pmatrix}$. Find a matrix, $S$ such that $S^{-1}AS = D$, a diagonal matrix.

Solving $\det(\lambda I - A) = 0$ yields the eigenvalues are 2 and 6 with 2 an eigenvalue of multiplicity two. Solving $(2I - A)x = 0$ to find the eigenvectors, you find that the eigenvectors are

$$a \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} + b \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

where $a, b$ are scalars. An eigenvector for $\lambda = 6$ is $\begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix}$. Let the matrix $S$ be

$$S = \begin{pmatrix} -2 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & -2 \end{pmatrix}$$

That is, the columns are the eigenvectors. Then

$$S^{-1} = \begin{pmatrix} -\frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \end{pmatrix}.$$
\[ S^{-1}AS = \begin{pmatrix} -\frac{1}{4} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 1 & 4 & -1 \\ -2 & -4 & 4 \end{pmatrix} \begin{pmatrix} -2 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & -2 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 6 \end{pmatrix}. \]

**Example 11.6.5** Here is a matrix. \[ A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix} \] Find \( A^{50} \).

Sometimes this sort of problem can be made easy by using diagonalization. In this case there are eigenvectors,
\[ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \]
the first two corresponding to \( \lambda = 1 \) and the last corresponding to \( \lambda = 2 \). Then let the eigenvectors be the columns of the matrix, \( S \). Thus
\[ S = \begin{pmatrix} 0 & -1 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \]

Then also
\[ S^{-1} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix} \]
and
\[ S^{-1}AS = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} = D \]

Now it follows
\[ A = SDS^{-1} = \begin{pmatrix} 0 & -1 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \]

Now note that \((SDS^{-1})^2 = SDS^{-1}SDS^{-1} = SD^2S^{-1} \) and
\[(SDS^{-1})^3 = SDS^{-1}SDS^{-1}SDS^{-1} = SD^3S^{-1}, \]
etc. In general, you can see that
\[ (SDS^{-1})^n = SD^nS^{-1} \]
In other words, \( A^n = SD^nS^{-1} \). Therefore,

\[
A^{50} = SD^{50}S^{-1} = \begin{pmatrix}
0 & -1 & -1 \\
0 & 1 & 0 \\
1 & 0 & 1
\end{pmatrix} \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 2
\end{pmatrix}^{50} = \begin{pmatrix}
1 & 1 & 1 \\
0 & 1 & 0 \\
-1 & -1 & 0
\end{pmatrix}.
\]

Now

\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 2
\end{pmatrix}^{50} = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 2^{50}
\end{pmatrix}.
\]

It follows

\[
A^{50} = \begin{pmatrix}
0 & -1 & -1 \\
0 & 1 & 0 \\
1 & 0 & 1
\end{pmatrix} \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 2^{50}
\end{pmatrix}^{50} \begin{pmatrix}
1 & 1 & 1 \\
0 & 1 & 0 \\
-1 & -1 & 0
\end{pmatrix} = \begin{pmatrix}
2^{50} & -1 + 2^{50} & 0 \\
0 & 1 & 0 \\
1 - 2^{50} & 1 - 2^{50} & 1
\end{pmatrix}.
\]

That isn’t too hard. However, this would have been horrendous if you had tried to multiply \( A^{50} \) by hand.

This technique of diagonalization is also important in solving the differential equations resulting from vibrations. Sometimes you have systems of differential equations and when you diagonalize an appropriate matrix, you “decouple” the equations. This is very nice. It makes hard problems trivial.

The above example is entirely typical. If \( A = SDS^{-1} \) then \( A^m = SD^mS^{-1} \) and it is easy to compute \( D^m \). More generally, you can define functions of the matrix using power series in this way. However, the real interesting case is when \( A \) is defective. This is much more interesting. You can always speak of things like \( \sin(A) \) for \( A \) an \( n \times n \) matrix. However, more interesting functions have no power series and you have to work harder for these. This is enough on this. One can go on and on.

### 11.7 The Matrix Exponential

When \( A \) is diagonalizable, one can easily define what is meant by \( e^A \). Here is how. You know

\[
S^{-1}AS = D
\]

where \( D \) is a diagonal matrix. You also know that if \( D \) is of the form

\[
\begin{pmatrix}
\lambda_1 & 0 \\
& \ddots \\
0 & \lambda_n
\end{pmatrix}
\]

then

\[
D^m = \begin{pmatrix}
\lambda_1^m & 0 \\
& \ddots \\
0 & \lambda_n^m
\end{pmatrix}
\]
and that
\[ A^m = SD^m S^{-1} \]
as shown above. Recall why this was.
\[ A = SDS^{-1} \]
and so
\[ A^m = \underbrace{SDS^{-1}SDS^{-1}\cdots SDS^{-1}}_{n \text{ times}} = SD^m S^{-1} \]

Now formally write the following power series for \( e^A \)
\[
e^A \equiv \sum_{k=0}^{\infty} \frac{A^k}{k!} = \sum_{k=0}^{\infty} \frac{SD^k S^{-1}}{k!} = S \sum_{k=0}^{\infty} \frac{D^k}{k!} S^{-1}
\]

If \( D \) is given above in \( \text{\textbullet\textbullet\textbullet} \), the above sum is of the form
\[
S \sum_{k=0}^{\infty} \left( \begin{array}{ccc}
\frac{1}{k!} \lambda_1^k & 0 & 0 \\
0 & \frac{1}{k!} \lambda_2^k & 0 \\
0 & 0 & \frac{1}{k!} \lambda_n^k
\end{array} \right) S^{-1} = S \left( \begin{array}{ccc}
\sum_{k=0}^{\infty} \frac{1}{k!} \lambda_1^k & 0 & 0 \\
0 & \sum_{k=0}^{\infty} \frac{1}{k!} \lambda_2^k & 0 \\
0 & 0 & \sum_{k=0}^{\infty} \frac{1}{k!} \lambda_n^k
\end{array} \right) S^{-1}
\]

and this last thing is the definition of what is meant by \( e^A \).

**Example 11.7.1** Let
\[
A = \begin{pmatrix}
2 & -1 & -1 \\
1 & 2 & 1 \\
-1 & 1 & 2
\end{pmatrix}
\]
Find \( e^A \).

The eigenvalues happen to be 1, 2, 3 and eigenvectors associated with these eigenvalues are
\[
\begin{pmatrix}
-1 \\
-1 \\
1
\end{pmatrix} \leftrightarrow 2, \quad \begin{pmatrix}
0 \\
-1 \\
1
\end{pmatrix} \leftrightarrow 1, \quad \begin{pmatrix}
-1 \\
0 \\
1
\end{pmatrix} \leftrightarrow 3
\]

Then let
\[
S = \begin{pmatrix}
-1 & 0 & -1 \\
-1 & -1 & 0 \\
1 & 1 & 1
\end{pmatrix}
\]
11.8. Migration Matrices

and so

\[ S^{-1} = \begin{pmatrix} -1 & -1 & -1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \]

and

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

Then the matrix exponential is

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

\[ \begin{pmatrix} e^2 & e^2 - e^3 & e^2 - e^3 \\ e^2 - e & e^2 & e^2 - e \\ -e^2 + e & -e^2 + e^3 & -e^2 + e + e^3 \end{pmatrix} \]

Isn’t that nice? You could also talk about \( \sin(A) \) or \( \cos(A) \) etc.

This matrix exponential is actually a useful idea when solving autonomous systems of first order linear differential equations. These are equations which are of the form

\[ x' = Ax \]

where \( x \) is a vector in \( \mathbb{R}^n \) or \( \mathbb{C}^n \) and \( A \) is an \( n \times n \) matrix. Then it turns out that the solution to the above system of equations is \( x(t) = e^{At}c \) where \( c \) is a constant vector.

11.8 Migration Matrices

There are applications of the eigenvalue problem which are of great importance and feature only one eigenvalue.

Consider the following table.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1/4</td>
<td>2/3</td>
</tr>
<tr>
<td>B</td>
<td>3/4</td>
<td>1/3</td>
</tr>
</tbody>
</table>

In this table, 1/4 is the probability that someone in location \( A \) ends up in \( A \) after a single unit of time. 2/3 is the probability that a person in location \( B \) ends up in location \( A \) after a single unit of time. 3/4 is the probability that a person in location \( A \) ends up in location \( B \) after a single unit of time and 1/3 is the probability that a person in location \( B \) ends up in location \( B \). Instead of the word probability, you could use the word “proportion” and the numbers would then represent the proportion of people in the various locations who end up in the other location after one unit of time. Thue 1/4 is the proportion of people in \( A \) who end up in \( A \), etc. Then this matrix is called a stochastic matrix, a Markov matrix or a Migration matrix. In the case the numbers are interpreted as probabilities, it is called a Markov or Stochastic matrix. In the case where they are proportions it is called a migration matrix.

Consider it as a migration matrix and suppose that initially there are 200 people in location \( A \) and 120 in location \( B \). You might wonder how many there would be in the two
locations after one unit of time. This is easy to figure out. Those in $A$ after one unit of time consist of those in $A$ who were in $A$ to begin with added to those in $A$ who started off in $B$. Thus

$$\text{# in } A = \frac{1}{4} (200) + \frac{2}{3} (120) = 130$$

$$\text{# in } B = \frac{3}{4} (200) + \frac{1}{3} (120) = 190.$$
After 100 units of time, you would have
\[
\begin{pmatrix}
 a_{100} \\
 b_{100}
\end{pmatrix} = \begin{pmatrix}
 \frac{1}{4} & \frac{3}{4} \\
 \frac{3}{4} & \frac{1}{4}
\end{pmatrix}^{100} \begin{pmatrix}
 a_0 \\
 b_0
\end{pmatrix} = \begin{pmatrix}
 .470588235 & .470588235 \\
 .529411765 & .529411765
\end{pmatrix} \begin{pmatrix}
 a_0 \\
 b_0
\end{pmatrix}
\]
You can’t detect any difference between these two answers. In general, if you wanted to know about how many would be in the two locations, you would need to take a limit. However, there is a better way.

More generally here is a definition.

**Definition 11.8.1** Let \( n \) locations be denoted by the numbers \( 1, 2, \ldots, n \). Also suppose it is the case that each year \( a_{ij} \) denotes the proportion of residents in location \( j \) which move to location \( i \). Also suppose no one escapes or emigrates from without these \( n \) locations.

This last assumption requires \( \sum_i a_{ij} = 1 \). 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 11.8.2** 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 11.8.3** 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,
\[
\det(A - \lambda I) = \det\left((A - \lambda I)^T\right) = \det(A^T - \lambda I)
\]
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, \( \sum_j a_{ij} = 1 \). 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,

\[
A^T \begin{pmatrix}
1 \\
\vdots \\
1
\end{pmatrix} = \begin{pmatrix}
\sum_j b_{ij} \\
\vdots \\
\sum_j b_{ij}
\end{pmatrix} = \begin{pmatrix}
1 \\
\vdots \\
1
\end{pmatrix}
\]

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 11.8.4** 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

\[
\begin{pmatrix}
.6 & 0 & .1 \\
.2 & .8 & 0 \\
.2 & .2 & .9
\end{pmatrix}^{10} \begin{pmatrix}
100 \\
200 \\
400
\end{pmatrix} = \begin{pmatrix}
115.085 & 829 & 22 \\
120.130 & 672 & 44 \\
464.783 & 498 & 34
\end{pmatrix}
\]

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

\[
x = \lim_{k \to \infty} A^k x_0 = \lim_{k \to \infty} A^{k+1} x_0 = A \lim_{k \to \infty} A^k x = A x,
\]

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 11.8.5** 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 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

\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix} - \begin{pmatrix}
.6 & 0 & .1 \\
.2 & .8 & 0 \\
.2 & .2 & .9
\end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
\]

The augmented matrix is

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

and its row reduced echelon form is

\[
\begin{pmatrix}
1 & 0 & -.25 & 0 \\
0 & 1 & -.25 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]

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

\[
\frac{1400}{3} \begin{pmatrix}
(1/4) \\
(1/4) \\
1
\end{pmatrix} = \begin{pmatrix}
116.666666666667 \\
116.666666666667 \\
466.666666666667
\end{pmatrix}
\]

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 11.8.6** Suppose a migration matrix is

\[
\begin{pmatrix}
\frac{1}{5} & \frac{1}{2} & \frac{1}{5} \\
\frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\
\frac{11}{20} & \frac{1}{4} & \frac{3}{10}
\end{pmatrix}
\]

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

$$
\begin{pmatrix}
4 & -1 & -1 & | & 0 \\
-\frac{1}{4} & \frac{3}{4} & -\frac{1}{2} & | & 0 \\
-\frac{11}{20} & -\frac{1}{4} & \frac{7}{10} & | & 0
\end{pmatrix}
$$

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 $\frac{16}{19}$ more in location 2 than in location 1. There will be $\frac{18}{19}$ 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.

### 11.9 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 11.9.1** 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

$$
det \left( \begin{pmatrix}
1 & 0 & 0 \\
0 & 2 & -1 \\
0 & 1 & 2
\end{pmatrix} - \lambda \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix} \right) = 0.
$$
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

$$(2 + i) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

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.

### 11.10 The Estimation Of Eigenvalues

There are ways to estimate the eigenvalues for matrices. The most famous is known as Gerschgorin’s theorem. This theorem gives a rough idea where the eigenvalues are just from looking at the matrix.
Theorem 11.10.1 Let \( A \) be an \( n \times n \) matrix. Consider the \( n \) Gerschgorin 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 Gerschgorin disc.

This theorem says to add up the absolute values of the entries of the \( i \)th row which are off the main diagonal and form the disc centered at \( a_{ii} \) having this radius. The union of these discs contains \( \sigma(A) \).

Proof: Suppose \( Ax = \lambda x \) where \( x \neq 0 \). Then for \( A = (a_{ij}) \), let \( |x_k| \geq |x_j| \) for all \( x_j \). Thus \( |x_k| \neq 0 \).

\[
\sum_{j \neq k} a_{kj} x_j = (\lambda - a_{kk}) x_k.
\]

Then

\[
|x_k| \sum_{j \neq k} |a_{kj}| \geq \sum_{j \neq k} |a_{kj}| |x_j| \geq \sum_{j \neq k} a_{kj} x_j = |\lambda - a_{ii}| |x_k|.
\]

Now dividing by \( |x_k| \), it follows \( \lambda \) is contained in the \( k \)th Gerschgorin disc.

Example 11.10.2 Here is a matrix. Estimate its eigenvalues.

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

According to Gerschgorin’s theorem the eigenvalues are contained in the disks

\[
\begin{align*}
D_1 &= \{ \lambda \in \mathbb{C} : |\lambda - 2| \leq 2 \}, \\
D_2 &= \{ \lambda \in \mathbb{C} : |\lambda - 5| \leq 3 \}, \\
D_3 &= \{ \lambda \in \mathbb{C} : |\lambda - 9| \leq 1 \}
\end{align*}
\]

It is important to observe that these disks are in the complex plane. In general this is the case. If you want to find eigenvalues they will be complex numbers.

So what are the values of the eigenvalues? In this case they are real. You can compute them by graphing the characteristic polynomial, \( \lambda^3 - 16\lambda^2 + 70\lambda - 66 \) and then zooming in on the zeros. If you do this you find the solution is \( \{ \lambda = 1.2953 \}, \{ \lambda = 5.5905 \}, \{ \lambda = 9.1142 \} \). Of course these are only approximations and so this information is useless for finding eigenvectors. However, in many applications, it is the size of the eigenvalues which is important and so these numerical values would be helpful for such applications.

Because of this example, you might think there is no real reason for Gerschgorin’s theorem. Why not just compute the characteristic equation and graph and zoom? This is fine up to
a point, but what if the matrix was huge? Then it might be hard to find the characteristic polynomial. Remember the difficulties in expanding a big matrix along a row or column. You would need a better way to come up with the characteristic polynomial. Also, what if the eigenvalue were complex? You don’t see these by following this procedure. However, Gerschgorin’s theorem will at least estimate them.

There are also more advanced versions of this theorem which depend on the theory of functions of a complex variable covering the case where the Gerschgorin disks are disjoint. In this case, you can assert each disk contains an eigenvalue. In fact, if \( k \) of the Gerschgorin disks are disjoint from the other disks then they contain \( k \) eigenvalues. To see this proved, see the linear algebra book on my web page. Don’t bother to look at it if you have not had a substantial course on complex analysis because it won’t make any sense. Math is not like comparative literature, history, or humanities. You can’t read the advanced topics until you have mastered the basic topics even if you are real smart.

### 11.11 Exercises With Answers

1. Find the eigenvectors and eigenvalues of the matrix, \( A = \begin{pmatrix} 8 & -3 & 1 \\ -2 & 7 & 1 \\ 0 & 0 & 10 \end{pmatrix} \). Determine whether the matrix is defective. If nondefective, diagonalize the matrix with an appropriate similarity transformation.

First you need to write the characteristic equation.

\[
\text{det} \left( \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 8 & -3 & 1 \\ -2 & 7 & 1 \\ 0 & 0 & 10 \end{pmatrix} \right) = \text{det} \begin{pmatrix} \lambda - 8 & 3 & -1 \\ 2 & \lambda - 7 & -1 \\ 0 & 0 & \lambda - 10 \end{pmatrix}
\]

\[
= \lambda^3 - 25\lambda^2 + 200\lambda - 500 = 0
\]

Next you need to find the solutions to this equation. Of course this is a real joy. If there are any rational zeros they are

\[
\pm \frac{\text{factor of 500}}{\text{factor of 1}}
\]

I hope to find a rational zero. If there are none, then I don’t know what to do at this point. This is a really lousy method for finding eigenvalues and eigenvectors. It only works if things work out well. Lets try 10. You can plug it in and see if it works or you can use synthetic division.

\[
\begin{array}{c|ccccc}
0 & 1 & -25 & 200 & -500 \\
10 & & 10 & -150 & 500 \\
\end{array}
\]

Yes, it appears 10 works and you can factor the polynomial as \((\lambda - 10)(\lambda^2 - 15\lambda + 50)\) which factors further to \((\lambda - 10)(\lambda - 5)(\lambda - 10)\) so you find the eigenvalues are 5, 10, and 10. It remains to find the eigenvectors. First find an eigenvector for \(\lambda = 5\). To do this, you find a vector which is sent to 0 by the matrix on the right in which you let \(\lambda = 5\). Thus the augmented matrix of the system of equations you need to
solve to get the eigenvector is

\[
\begin{pmatrix}
5 - 8 & 3 & -1 & | & 0 \\
2 & 5 - 7 & -1 & | & 0 \\
0 & 0 & 5 - 10 & | & 0
\end{pmatrix}
\]

Now the row reduced echelon form is

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

and so you need \( x = y \) and \( z = 0 \). An eigenvector is \((1, 1, 0)^T\). Now you have the glorious opportunity to solve for the eigenvectors associated with \( \lambda = 10 \). You do it the same way. The augmented matrix for the system of equations you solve to find the eigenvectors is

\[
\begin{pmatrix}
10 - 8 & 3 & -1 & | & 0 \\
2 & 10 - 7 & -1 & | & 0 \\
0 & 0 & 10 - 10 & | & 0
\end{pmatrix}
\]

The row reduced echelon form is

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

and so you need \( x = -\frac{3}{2}y + \frac{1}{2}z \). It follows the eigenvectors for \( \lambda = 10 \) are

\[
\left( \frac{3}{2}y + \frac{1}{2}z, y, z \right)^T
\]

where \( x, y \in \mathbb{R}, \) not both equal to zero. Why? Let \( y = 2 \) and \( z = 0 \). This gives the vector,

\[
(-3, 2, 0)^T
\]
as one of the eigenvectors. You could also let \( y = 0 \) and \( z = 2 \) to obtain another eigenvector,

\[
(1, 0, 2)^T.
\]

If there exists a basis of eigenvectors, then the matrix is nondefective and as discussed above, the matrix can be diagonalized by considering \( S^{-1}AS \) where the columns of \( S \) are the eigenvectors. In this case, I have found three eigenvectors and so it remains to determine whether these form a basis. Remember how to do this. You let them be the columns of a matrix and then find the rank of this matrix. If it is three, then they are a basis because they are linearly independent and the vectors are in \( \mathbb{R}^3 \). This is equivalent to the following matrix has an inverse.

\[
\begin{pmatrix}
1 & -3 & 1 \\
1 & 2 & 0 \\
1 & 0 & 2
\end{pmatrix}
\]
Then to diagonalize
\[
\begin{pmatrix}
\frac{2}{5} & \frac{3}{5} & -\frac{1}{5} \\
-\frac{1}{5} & \frac{2}{5} & \frac{1}{10} \\
0 & 0 & \frac{1}{2}
\end{pmatrix}
\begin{pmatrix}
8 & -3 & 1 \\
-2 & 7 & 1 \\
0 & 0 & 10
\end{pmatrix}
\begin{pmatrix}
1 & -3 & 1 \\
1 & 2 & 0 \\
0 & 0 & 2
\end{pmatrix} =
\begin{pmatrix}
5 & 0 & 0 \\
0 & 10 & 0 \\
0 & 0 & 10
\end{pmatrix}
\]

Isn’t this stuff marvelous! You can know this matrix is nondefective at the point when you find the eigenvectors for the repeated eigenvalue. This eigenvalue was repeated with multiplicity 2 and there were two parameters, \(y\) and \(z\) in the description of the eigenvectors. Therefore, the matrix is nondefective. Also note that there is no uniqueness for the similarity transformation.

2. Now consider the matrix, \(
\begin{pmatrix}
2 & 1 & 0 \\
0 & 1 & 0 \\
-1 & 0 & 1
\end{pmatrix}
\). Find its eigenvectors and eigenvalues and determine whether it is defective.

The characteristic equation is
\[
\det \left( \lambda \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix} - \begin{pmatrix}
2 & 1 & 0 \\
0 & 1 & 0 \\
-1 & 0 & 1
\end{pmatrix} \right) = 0
\]

thus the characteristic equation is
\[
(\lambda - 2)(\lambda - 1)^2 = 0.
\]

The zeros are 1, 1, 2. Let's find the eigenvectors for \(\lambda = 1\). The augmented matrix for the system you need to solve is
\[
\begin{pmatrix}
-1 & -1 & 0 & | & 0 \\
0 & 0 & 0 & | & 0 \\
1 & 0 & 0 & | & 0
\end{pmatrix}
\]

The row reduced echelon form is
\[
\begin{pmatrix}
1 & 0 & 0 & | & 0 \\
0 & 1 & 0 & | & 0 \\
0 & 0 & 0 & | & 0
\end{pmatrix}
\]

Then you find \(x = y = 0\) and there is no restriction on \(z\). Thus the eigenvectors are of the form
\[
(0, 0, z)^T, \ z \in \mathbb{R}.
\]
The eigenvalue had multiplicity 2 but the eigenvectors depend on only one parameter. Therefore, the matrix is defective and cannot be diagonalized. The other eigenvector comes from row reducing the following

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

The row reduced echelon form is

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

Therefore the eigenvectors are of the form

\[(x, 0, -x)^T.
\]

One such eigenvector is \((1, 0, -1)^T\).

3. 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

\[
x^*x = \sum_k x_k^* x_k = \sum_k |x_k|^2 = |x|^2.
\]

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 \(Mx = \lambda x\)

\[
\overline{\lambda} x^* x = (\lambda x)^* x = (Mx)^* x = x^* M^* x
\]

and so \(\lambda = \overline{\lambda}\) showing that \(\lambda\) must be real.

4. 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,

\[
0 = p(\lambda) = \overline{p(\overline{\lambda})} = p(\overline{\lambda})
\]

showing that \(\overline{\lambda}\) is also an eigenvalue.
5. 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.

6. Here is a matrix. \( A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & -1 & 0 \\ 0 & -2 & 1 \end{pmatrix} \). Find a formula for \( A^n \) where \( n \) is an integer.

First you find the eigenvectors and eigenvalues. \( \begin{pmatrix} 1 & 2 & 0 \\ 0 & -1 & 0 \\ 0 & -2 & 1 \end{pmatrix} \), eigenvectors:

\[
\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \leftrightarrow 1, \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} \leftrightarrow -1.
\]

The matrix, \( S \) used to diagonalize the matrix is obtained by letting these vectors be the columns of \( S \). Then \( S^{-1} \) is given by

\[
S^{-1} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix}
\]
Then $S^{-1}AS$ equals
\[
\begin{pmatrix}
1 & 1 & 0 \\
0 & -1 & 1 \\
0 & 1 & 0
\end{pmatrix}
\begin{pmatrix}
1 & 2 & 0 \\
0 & -1 & 0 \\
0 & -2 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 0 & -1 \\
0 & 0 & 1 \\
0 & 1 & 1
\end{pmatrix}
= \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & -1
\end{pmatrix} \equiv D
\]
Then $A = SDS^{-1}$ and $A^n = SD^nS^{-1}$. Now it is easy to find $D^n$.

\[
D^n = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & (-1)^n
\end{pmatrix}
\]

Therefore,
\[
A^n = \begin{pmatrix}
1 & 0 & -1 \\
0 & 0 & 1 \\
0 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & (-1)^n
\end{pmatrix}
\begin{pmatrix}
1 & 1 & 0 \\
0 & 0 & (-1)^n \\
0 & -1 & 1
\end{pmatrix}
= \begin{pmatrix}
1 & 1 & 0 \\
0 & 0 & (-1)^n \\
0 & -1 & 1 + (-1)^n
\end{pmatrix}
\]

7. Suppose the eigenvalues of $A$ are $\lambda_1, \ldots, \lambda_n$ and that $A$ is nondefective. Show that
\[
e^{At} = S \begin{pmatrix}
e^{\lambda_1 t} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & e^{\lambda_n t}
\end{pmatrix} S^{-1}
\]
where $S$ is the matrix which satisfies $S^{-1}AS = D$.

The diagonal matrix, $D$ has the same characteristic equation as $A$ why? and so it has the same eigenvalues. However the eigenvalues of $D$ are the diagonal entries and so the diagonal entries of $D$ are the eigenvalues of $A$. Now
\[
S^{-1}tAS = tD
\]
and
\[
(tD)^n = \begin{pmatrix}
(\lambda_1 t)^n & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & (\lambda_n t)^n
\end{pmatrix}
\]

Therefore,
\[
\sum_{n=0}^{\infty} \frac{1}{n!} (tD)^n = \sum_{n=0}^{\infty} \frac{(S^{-1}tAS)^n}{n!} = S^{-1} \sum_{n=0}^{\infty} \frac{(tA)^n}{n!} S.
\]
Now the left side equals
\[
\sum_{n=0}^{\infty} \frac{1}{n!} (tD)^n = \sum_{n=0}^{\infty} \frac{1}{n!} \begin{pmatrix} (\lambda_1 t)^n & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & (\lambda_n t)^n \end{pmatrix}
\]
\[
= \begin{pmatrix} \sum_{n=0}^{\infty} \frac{(\lambda_1 t)^n}{n!} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_{n=0}^{\infty} \frac{(\lambda_n t)^n}{n!} \end{pmatrix}
\]
\[
= \begin{pmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{pmatrix}
\]

Therefore,
\[
e^{tA} = \sum_{n=0}^{\infty} \frac{(tA)^n}{n!} = S \begin{pmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{pmatrix} S^{-1}.
\]

8. Show that if \( A \) is similar to \( B \) then \( A^T \) is similar to \( B^T \).

This is easy. \( A = S^{-1} BS \) and so \( A^T = S^T B^T (S^{-1})^T = S^T B^T (S^T)^{-1} \).

9. Suppose \( A^m = 0 \) for some \( m \) a positive integer. Show that if \( A \) is diagonalizable, then \( A = 0 \).

Since \( A^m = 0 \) suppose \( S^{-1} AS = D \). Then raising to the \( m^{th} \) power, \( D^m = S^{-1} A^m S = 0 \). Therefore, \( D = 0 \). But then \( A = S0S^{-1} = 0 \).

10. 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,

\[ A \mathbf{v} = \overline{A \mathbf{v}} = \lambda \mathbf{v}. \]

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

\[ \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}. \]
Chapter 12
Curves In Space

12.1 Limits Of A Vector Valued Function Of One Variable

Quiz

1. Find the determinant of the matrix,

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

2. Find all eigenspaces and eigenvalues for the matrix,

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

3. Find the eigenspace for the eigenvalue \( \lambda = 2 \) for the matrix

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

4. If \( A \) is a real matrix with a complex eigenvalue \( \lambda = a + ib \) where \( b \neq 0 \), show that if \( v \) is an eigenvector, then

\[ v = x + iy \]

where \( y \neq 0 \).

A vector valued function is just one which has vector values. For example, consider

\[ (\cos t, t^2, t + 1) \]

where \( t \in [0, 2] \). Each value of \( t \) corresponds to a point in \( \mathbb{R}^3 \) whose coordinates are as given. Thus when \( t = 0 \), the point in \( \mathbb{R}^3 \) is \((1,0,1)\) and when \( t = \pi/2 \) the point is
\( \left( 0, \left( \frac{\pi}{2} \right)^2, \frac{\pi}{2} + 1 \right) \), etc. Often \( t \) will be considered as time. Thus, in this case, the vector valued function gives the coordinates of a point which is moving in three dimensions as a function of time. Imagine a fly buzzing around the room for example. Let the origin be a corner of the room and consider the position vector of the fly. This position vector could be described by a vector valued function of the form \((x(t), y(t), z(t))\) where \( t \) is in some interval. Here \( x(t) \) is the \( x \) coordinate of the fly, \( y(t) \) the \( y \) coordinate, and \( z(t) \), the \( z \) coordinate corresponding to a given time. Later the physical significance of all this will be discussed more. For right now, \( t \) will just be in some interval and general vector valued functions will be considered.

**Definition 12.1.1** Let \( x(t) = (x_1(t), \ldots, x_n(t)) \) for \( t \in [a,b] \) be a vector valued function. The curve **parametrized** by this vector valued function is the set of points in \( \mathbb{R}^n \) which are obtained by letting \( t \) vary over the interval, \( [a,b] \). The vector valued function is also called a **parametrization** of this curve. The variable, \( t \) is called a **parameter**. More generally, if \( x(t) = (x_1(t), \ldots, x_n(t)) \) is given where each \( x_i(t) \) is a formula the **domain** of \( x \) is defined to be the set where each of the \( x_i(t) \) is defined. It is denoted by \( D(x) \).

**Example 12.1.2** Let \( x(t) = (\frac{1}{t}, \sqrt{1-t^2}, \sin(t)) \) find the domain of \( x \).

You need each function to make sense. Thus you must have \(-1 \leq t \leq 1 \) and \( t \neq 0 \). The domain is \([-1,0) \cup (0,1]\).

In useful situations the domain will typically be an interval.

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 12.1.3** 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.

\[
\lim_{t \to \infty} f(t) = L
\]
12.2. THE DERIVATIVE AND INTEGRAL

if for every \( \varepsilon > 0 \) there exists \( l \) such that whenever \( t > l \),
\[
|f(t) - L| < \varepsilon
\] (12.1)
and
\[
\lim_{t \to -\infty} f(t) = L
\]
if for every \( \varepsilon > 0 \) there exists \( l \) such that whenever \( t < l \), (12.2)
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 \( p > 1 \).

Observation 12.1.4 Let \( f(t) = (f_1(t), \cdots, f_n(t)) \) and let \( L = (L_1, \cdots, L_n) \). Then \( \lim_{t \to a} f(t) = L \) if and only if \( \lim_{t \to a} f_k(t) = L_k \) for each \( k \).

Example 12.1.5 Let \( f(t) = (\cos t, \sin t, t^2 + 1, \ln(t)) \). Find \( \lim_{t \to \pi/2} f(t) \).

Using the above observation, this limit equals
\[
\left( \lim_{t \to \pi/2} \cos t, \lim_{t \to \pi/2} \sin t, \lim_{t \to \pi/2} (t^2 + 1), \lim_{t \to \pi/2} \ln(t) \right)
= \left( 0, 1, \frac{\pi^2}{4} + 1, \ln\left(\frac{\pi}{2}\right) \right).
\]

Example 12.1.6 Let \( f(t) = (\sin t, t^2, t + 1) \). Find \( \lim_{t \to 0} f(t) \).

Recall that \( \lim_{t \to 0} \frac{\sin t}{t} = 1 \). Then using the above observation, \( \lim_{t \to 0} f(t) = (1, 0, 1) \).

12.2 The Derivative And Integral

The following definition is on the derivative and integral of a vector valued function of one variable.

Definition 12.2.1 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 \( f(t) = (f_1(t), \cdots, f_p(t)) \) and \( \int_a^b f_i(t) \, dt \) exists for each \( i = 1, \cdots, p \), then
\[
\int_a^b f(t) \, dt \text{ is defined as the vector,}
\]
\[
\left( \int_a^b f_1(t) \, dt, \cdots, \int_a^b f_p(t) \, dt \right)
\]
This is what is meant by saying \( f \in R([a,b]) \). In other words, \( f \) is Riemann integrable. That is you can take the integral.

It is easier to write \( f \in R([a,b]) \) than to write \( f \) is Riemann integrable. Thus, if you see \( f \in R([a,b]) \), think: \( \int_a^b f(x) \, dx \) exists.

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 12.2.2 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
\[
|f (t + h) - f (t)| < |h| + |f'(t)| |h| .
\]
Now letting \( \delta = \min \left( \delta_1, \frac{\varepsilon}{1 + |f'(t)|} \right) \) 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 12.2.3 If \( f \in R ([a, b]) \) and if \( f \) is continuous at \( t \in (a, b) \), then
\[
\frac{d}{dt} \left( \int_{a}^{t} f (s) \, ds \right) = f (t) .
\]

Proof: Say \( f (t) = (f_1 (t), \cdots, f_p (t)) \). Then it follows
\[
1 \int_{a}^{t+h} f (s) \, ds - 1 \int_{a}^{t} f (s) \, ds = \left( \frac{1}{h} \int_{t}^{t+h} f_1 (s) \, ds, \cdots, \frac{1}{h} \int_{t}^{t+h} f_p (s) \, ds \right)
\]
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.

Example 12.2.4 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,
\[
\lim_{h \to 0} \frac{f (x + h) - f (x)}{h} = \lim_{h \to 0} 0 = 0
\]

Example 12.2.5 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 \( f' (t) = (a, b) \).
12.2. THE DERIVATIVE AND INTEGRAL

12.2.1 Arc Length

$C$ 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. $x'_i$ 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 $p(t) \equiv (x_1(t), \ldots, x_n(t))$, $t \rightarrow p(t)$ is one to one on $(a, b)$.
4. $|p'(t)| \equiv \left( \sum_{i=1}^{n} |x'_i(t)|^2 \right)^{1/2} \neq 0$ for all $t \in [a, b]$.
5. $C = \cup \{ (x_1(t), \ldots, x_n(t)) : t \in [a, b] \}$.

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 **parametrization** 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 parametrization $(x_1, \ldots, x_n)$. Forming a partition of $[a, b]$, $a = t_0 < \cdots < t_n = b$ and letting $p_i = (x_1(t_i), \ldots, x_n(t_i))$, 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.

Now consider what happens when the partition is refined by including more points. You can see from the following picture that the polygonal approximation would appear to be even better and that as more points are added in the partition, the sum of the lengths of the line segments seems to get close to something which deserves to be defined as the length of the curve, $C$. The following picture illustrates what is meant by this.
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 parametrization 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 parametrization 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 parametrization. It is proved in Section 14.2.2 which starts on Page 297.

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 parametrization for the line segment joining these two points is
\[
f_i(t) \equiv tp_i + (1-t)q_i, \quad t \in [0,1].
\]
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.

**Definition 12.2.6** A curve \( C \) is piecewise smooth if there exist points on this curve, \( p_0, p_1, \ldots, p_n \) such that, denoting \( C_{p_{k-1}p_k} \) the part of the curve joining \( p_{k-1} \) and \( p_k \), it follows \( C_{p_{k-1}p_k} \) is a smooth curve and \( \bigcup_{k=1}^{n} C_{p_{k-1}p_k} = C \). In other words, it is piecewise smooth if it consists of a finite number of smooth curves linked together.

To find the length of a piecewise smooth curve, just sum the lengths of the smooth pieces described above.

**Example 12.2.7** The parametrization for a smooth curve is \( r(t) = (t, 2t^2, t^3) \) for \( t \in [0,1] \). Find the length of this curve.

From the above, the length is
\[
\int_{0}^{1} |r'(t)| \, dt = \int_{0}^{1} \sqrt{(4t)^2 + (2t)^2 + t^2} \, dt = \frac{1}{2} \sqrt{21} + \frac{1}{20} \sqrt{5} \ln \left( 2 \sqrt{5} + \sqrt{21} \right).
\]
You need to use a trig substitution of some sort to do this integral but it is routine. Of course, if you can’t find an antiderivative, then you solve it numerically.
Example 12.2.8 The parametrization for a smooth curve is \( r(t) = (t, t^3, t^2) \) for \( t \in [0, 1] \).
Find the length of this curve.

The length is

\[
\int_0^1 \sqrt{1 + (3t^2)^2 + (2t)^2} \, dt = \int_0^1 \sqrt{1 + 9t^4 + 4t^2} \, dt
\]

and I have no clue how to find an antiderivative for this. Therefore, I find the integral numerically.

\[
\int_0^1 \sqrt{1 + 9t^4 + 4t^2} \, dt = 1.863
\]

This is all right to do. Numerical methods are allowed and sometimes that is all you can get.

12.2.2 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 \( r \). It is assumed \( r'(t) \) exists, is nonzero, and \( r \) is one to one on some \((a,b)\) such that \( t \in (a,b) \).

![Diagram](image)

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

\[
\frac{r(t+h) - r(t)}{|r(t+h) - r(t)|} \equiv T_h.
\]

Thus \( T_h \to T \), a unit tangent vector to the curve at the point \( r(t) \). Therefore,

\[
r'(t) \equiv \lim_{h \to 0} \frac{r(t+h) - r(t)}{h} = \lim_{h \to 0} \frac{|r(t+h) - r(t)|}{h} \frac{r(t+h) - r(t)}{|r(t+h) - r(t)|} = \lim_{h \to 0} \frac{|r(t+h) - r(t)|}{h} T_h = |r'(t)| T.
\]

Now what is the significance of \( |r'(t)| \) in the special case that \( t \) is time? Recall the length of the curve from \( r(t) \) to \( r(t + h) \) is

\[
\int_t^{t+h} |r'(s)| \, ds
\]

This is the distance travelled along the curve during the time interval \([t, t+h]\). Therefore,

\[
\frac{1}{h} \int_t^{t+h} |r'(s)| \, ds
\]
is the average speed during this time interval. Letting $h \to 0$, the limit in the above is just $|r'(t)|$ by the fundamental theorem of calculus. Therefore, the significance of $|r'(t)|$ is that it is the instantaneous speed of the object at time $t$. Thus $|r'(t)| \mathbf{T}$ represents the speed times a unit direction vector, $\mathbf{T}$ 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.

How do you go about computing $r'(t)$? Letting $r(t) = (r_1(t), \ldots, r_q(t))$, the expression

$$
\frac{r(t_0 + h) - r(t_0)}{h}
$$

is equal to

$$
\left( \frac{r_1(t_0 + h) - r_1(t_0)}{h}, \ldots, \frac{r_q(t_0 + h) - r_q(t_0)}{h} \right).
$$

Then as $h$ converges to 0, (12.2) converges to

$$
v \equiv (v_1, \ldots, v_q)
$$

where $v_k = r'_k(t)$. This by Observation [12.1.4], which says that the term in (12.2) gets close to a vector, $v$ if and only if all the coordinate functions of the term in (12.2) get close to the corresponding coordinate functions of $v$.

In the case where $t$ is time, this simply says the velocity vector equals the vector whose components are the derivatives of the components of the displacement vector, $r(t)$.

In any case, the vector, $\mathbf{T}$ determines a direction vector which is tangent to the curve at the point, $r(t)$ and so it is possible to find parametric equations for the line tangent to the curve at various points.

**Example 12.2.9** Let $r(t) = (\sin t, t^2, t + 1)$ for $t \in [0, 5]$. Find a tangent line to the curve parametrized by $\mathbf{r}$ at the point $\mathbf{r}(2)$.

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

$$(\sin 2, 4, 3) + (t \cos 2, 4, 1) = (x, y, z).$$

**Example 12.2.10** Let $r(t) = (\sin t, t^2, t + 1)$ for $t \in [0, 5]$. Find the velocity vector when $t = 1$.

From the above discussion, this is simply $r'(1) = (\cos 1, 2, 1)$.

### 12.2.3 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 12.2.11** Let $a, b \in \mathbb{R}$ and suppose $f'(t)$ and $g'(t)$ exist. Then the following formulas are obtained.

$$
(a f + b g)'(t) = a f'(t) + b g'(t).
$$

(12.3)

$$
(f \cdot g)'(t) = f'(t) \cdot g(t) + f(t) \cdot g'(t).
$$

(12.4)

If $f, g$ have values in $\mathbb{R}^3$, then

$$
(f \times g)'(t) = f'(t) \times g(t) + f(t) \times g'(t).
$$

(12.5)

The formulas, (12.3), (12.4), and (12.5) are referred to as the product rule.
Proof: The first formula is left for you to prove. Consider the second. Let
\[
\lim_{h \to 0} \frac{f \cdot g(t + h) - f(t) \cdot g(t)}{h}
\]
\[
= \lim_{h \to 0} \frac{f(t + h) \cdot g(t + h) - f(t) \cdot g(t)}{h}
\]
\[
= \lim_{h \to 0} \left( f(t + h) \cdot \left( \frac{g(t + h) - g(t)}{h} \right) + \frac{f(t + h) - f(t)}{h} \cdot g(t) \right)
\]
\[
= \lim_{h \to 0} \sum_{k=1}^{n} f_k(t + h) \left( \frac{g_k(t + h) - g_k(t)}{h} \right) + \sum_{k=1}^{n} \left( \frac{f_k(t + h) - f_k(t)}{h} \right) g_k(t)
\]
\[
= \sum_{k=1}^{n} f_k(t) g_k(t) + \sum_{k=1}^{n} f_k(t) g_k(t)
\]
\[
= f'(t) \cdot g(t) + f(t) \cdot g'(t).
\]

Formula (12.3) 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 266. You can also see this is true by using the distributive law of the cross product.

\[
f(t + h) \times g(t + h) - f(t) \times g(t)
\]
\[
= f(t + h) \times g(t + h) - f(t + h) \times g(t) + f(t + h) \times g(t) - f(t) \times g(t)
\]

and so

\[
\frac{1}{h} (f(t + h) \times g(t + h) - f(t) \times g(t))
\]
\[
= f(t + h) \times \left( \frac{g(t + h) - g(t)}{h} \right) + \left( \frac{f(t + h) - f(t)}{h} \right) \times g(t)
\]

Now assuming the cross product is continuous, (This is obvious from either the component or the geometric description of the cross product.) you can take a limit in the above as \(h \to 0\) and obtain

\[
f(t) \times g'(t) + f'(t) \times g(t).
\]

It is exactly like the product rule for scalar valued functions except you need to be very careful about the order in which things are multiplied because the cross product is not commutative.

Example 12.2.12 Let

\[
r(t) = (t^2, \sin t, \cos t)
\]

and let \(p(t) = (t, \ln (t + 1), 2t)\). Find \((r(t) \times p(t))'\).

From (12.3) this equals\(2t, \cos t, -\sin t\) \times \(t, \ln (t + 1), 2t\) + \((t^2, \sin t, \cos t) \times (1, \frac{1}{t + 1}, 2)\)

\[
= (2 \cos t t + \sin t \ln (t + 1), -\sin t t - 4t^2, 2t \ln (t + 1) - (\cos t) t)
\]
\[
+ \left( 2 \sin t - \frac{\cos t}{t + 1}, \cos t - 2t^2, \frac{t^2}{t + 1} - \sin t \right)
\]
\[
= (2 \cos t t + \sin t \ln (t + 1) + 2 \sin t - \frac{\cos t}{t + 1}, -\sin t t - 6t^2 + \cos t,
\]
\[
2t \ln (t + 1) - (\cos t) t + \frac{t^2}{t + 1} - \sin t)
\]
Example 12.2.13  Let \( \mathbf{r}(t) = (t^2, \sin t, \cos t) \) Find \( \int_0^\pi \mathbf{r}(t) \, dt \).

This equals \( (\int_0^\pi t^2 \, dt, \int_0^\pi \sin t \, dt, \int_0^\pi \cos t \, dt) = (\frac{4}{3} \pi^3, 2, 0) \) .

Example 12.2.14  An object has position \( \mathbf{r}(t) = \left( t^3, \frac{t}{1 + t^2}, \sqrt{t^2 + 2} \right) \) 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.

\[
\mathbf{r}'(t) = \left( 3t^2, \frac{1}{1 + t^2} \left( \frac{1}{2} \right)^{-1/2}, \frac{1}{\sqrt{t^2 + 2}} \right)
\]

When \( t = 1 \), the velocity is

\[
\mathbf{r}'(1) = \left( 3, \frac{1}{4}, \frac{1}{\sqrt{3}} \right) \text{ kilometers per hour.}
\]

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 12.2.15  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 12.2.16  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 \( P_0 \) and the point where your sound is to be reflected is \( P_1 \). Consider a plane which contains the two points and let \( \mathbf{r}(t) \) denote a parametrization 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 \( P_0 - \mathbf{r}(t) \) and \( -\mathbf{r}'(t) \) equals the angle between \( P_1 - \mathbf{r}(t) \) and \( \mathbf{r}'(t) \). Draw a picture to see this. Therefore,

\[
\frac{(P_0 - r(t)) \cdot (-r'(t))}{|P_0 - r(t)||r'(t)|} = \frac{(P_1 - r(t)) \cdot (r'(t))}{|P_1 - r(t)||r'(t)|}.
\]
This reduces to
\[
\frac{(r(t) - P_0) \cdot (-r'(t))}{|r(t) - P_0|} = \frac{(r(t) - P_1) \cdot (r'(t))}{|r(t) - P_1|}
\] (12.6)

Now
\[
\frac{(r(t) - P_1) \cdot (r'(t))}{|r(t) - P_1|} = \frac{d}{dt} |r(t) - P_1|
\]
and a similar formula holds for \( P_1 \) replaced with \( P_0 \). This is because
\[
|r(t) - P_1| = \sqrt{(r(t) - P_1) \cdot (r(t) - P_1)}
\]
and so using the chain rule and product rule,
\[
\frac{d}{dt} |r(t) - P_1| = \frac{1}{2} ((r(t) - P_1) \cdot (r(t) - P_1))^{-1/2} 2 ((r(t) - P_1) \cdot r'(t))
\]
\[
= \frac{(r(t) - P_1) \cdot (r'(t))}{|r(t) - P_1|}.
\]
Therefore, from (12.6)
\[
\frac{d}{dt} (|r(t) - P_1|) + \frac{d}{dt} (|r(t) - P_0|) = 0
\]
showing that \( |r(t) - P_1| + |r(t) - P_0| = C \) for some constant, \( C \). This implies the curve of intersection of the plane with the room is an ellipse having \( P_0 \) and \( P_1 \) as the foci.

### 12.2.4 Leibniz’s Notation

Leibniz’s notation also generalizes routinely. For example, \( \frac{dy}{dt} = y'(t) \) with other similar notations holding.

### 12.2.5 Motion Of A Projectile (Idealized)

As we all know from beginning physics classes, the earth is flat and does not rotate. There is also no air resistance.

**Example 12.2.17** Under the above simplifying assumptions, suppose a rock is thrown from a height of \( a \) feet above the ground with a speed of \( v \) feet per second at an angle of \( \theta \) with the horizontal. Describe the motion of the rock. The following is a picture which is descriptive of the initial condition of this rock. Shown is the initial velocity \( v_0 \) as well as the initial position.

![Diagram](image)

Assume the motion is in the \( xy \) plane with the \( x \) axis corresponding to the ground. Let \( r(t) = (x(t), y(t)) \) be the motion of the rock.

\[
r''(t) = -32j
\]
Therefore,
\[ r'(t) = -32jt + c \]
where \( c \) is a constant vector. When \( t = 0 \), \( r'(0) = v_0 \) which is shown in the above picture. Thus \( c = v_0 \) and so
\[ r'(t) = -32jt + v_\cos(\theta)i + v_\sin(\theta)j \]
Then
\[ r(t) = -16t^2j + v_\cos(\theta)t i + v_\sin(\theta)t j + d \]
where \( d \) is another constant vector. It must equal \( aj \). Therefore,
\[ r(t) = (v_\cos(\theta)t, v_\sin(\theta)t - 16t^2 + a) \]

Example 12.2.18 In the above example, find where the rock hits the ground. Also find how high it goes. If \( a = 0 \), find \( \theta \) such that the range will be as large as possible.

The rock hits the ground when the second component equals 0. This occurs when
\[ v_\sin(\theta)t - 16t^2 + a = 0 \]
There are two solutions but you want the one for which \( t > 0 \) because time is increasing. Thus
\[ \frac{1}{32}v_\sin(\theta) + \frac{1}{32}\sqrt{(v^2\sin^2(\theta) + 64a)} \]
Then the \( x \) coordinate of where it hits the ground is
\[ v_\cos(\theta)\left(\frac{1}{32}v_\sin(\theta) + \frac{1}{32}\sqrt{(v^2\sin^2(\theta) + 64a)}\right) \quad (12.7) \]
To find when it is highest, you only have to maximize the function
\[ v_\sin(\theta)t - 16t^2 + a \]
You know how to do this. You take the derivative and set it equal to 0 to find the time at which the rock is highest. Thus
\[ v_\sin(\theta) - 32t = 0 \]
and so this time is
\[ t = \frac{v_\sin(\theta)}{32} \]
Now you plug this in to the second component to find the maximum height.
\[ v_\sin(\theta)\left(\frac{1}{32}v_\sin(\theta) - 16\left(\frac{v_\sin(\theta)}{32}\right)^2 + a \right) = \frac{1}{64}v^2\sin^2(\theta) + a \]
In case \( a = 0 \), what angle gives the maximum range? Consider the formula for the range in \[12.7\]. How can you make it as large as possible in case \( a = 0 \)? You want to maximize
\[ v_\cos(\theta)\left(\frac{1}{32}v_\sin(\theta) + \frac{1}{32}\sqrt{v^2\sin^2(\theta)}\right) = \frac{1}{16}v^2\cos\theta\sin\theta = \frac{1}{32}v^2\cos\theta\sin\theta = \frac{1}{32}v^2\sin(2\theta) \]
Obviously this is as large as possible when \( 2\theta = \pi/2 \). Thus the maximum range in this case occurs when \( \theta = \pi/4 \).
12.3 Exercises With Answers

1. Find the following limits if possible

   (a) \( \lim_{x \to 0^+} \left( \frac{|x|}{x}, \sin \frac{2x}{x}, \tan \frac{x}{x} \right) = (1, 2, 1) \)

   (b) \( \lim_{x \to 0^+} \left( \frac{x}{|x|}, \cos x, e^{2x} \right) = (1, 1, 1) \)

   (c) \( \lim_{x \to -4} \left( \frac{x^2 - 16}{x+4}, x - 7, \tan \frac{7x}{5x} \right) = (0, -3, \frac{7}{5}) \)

2. Let \( \mathbf{r}(t) = \left( 4 + (t - 1)^2, \sqrt{t^2 + 1} (t - 1)^3, \frac{(t-1)^3}{t^2} \right) \) 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.

   You need to differentiate this. \( \mathbf{r}'(t) = \left( 2(t-1), (t-1)^2 \frac{4t^2-t+3}{\sqrt{t^2+1}}, -(t-1)^2 \frac{2t-5}{t^6} \right) \).

   Now you need to find the value(s) of \( t \) where \( \mathbf{r}'(t) = 0 \).

3. Let \( \mathbf{r}(t) = (\sin t, t^2, 2t+1) \) for \( t \in [0, 4] \). Find a tangent line to the curve parametrized by \( \mathbf{r} \) at the point \( \mathbf{r}(2) \).

   \( \mathbf{r}'(t) = (\cos t, 2t, 2) \). When \( t = 2 \), the point on the curve is \((\sin 2, 4, 5)\). A direction vector is \( \mathbf{r}'(2) \) and so a tangent line is \( \mathbf{r}(t) = (\sin 2, 4, 5) + t(\cos 2, 4, 2) \).

4. Let \( \mathbf{r}(t) = (\sin t, \cos (t^2), t+1) \) for \( t \in [0, 5] \). Find the velocity when \( t = 3 \).

   \( \mathbf{r}'(t) = (\cos t, -2t \sin (t^2), 1) \). The velocity when \( t = 3 \) is just \( \mathbf{r}'(3) = (\cos 3, -6 \sin (9), 1) \).

5. 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 \( (\mathbf{r}(t) \times \mathbf{s}(t) \cdot \mathbf{p}(t))^t \).

   From the product rules for the cross and dot product, this equals

   \[ (\mathbf{r}(t) \times \mathbf{s}(t))^t \cdot \mathbf{p}(t) + (\mathbf{r}(t) \times \mathbf{s}(t)) \cdot \mathbf{p}'(t) = \mathbf{r}'(t) \times \mathbf{s}(t) \cdot \mathbf{p}(t) + \mathbf{r}(t) \times \mathbf{s}'(t) \cdot \mathbf{p}(t) + \mathbf{r}(t) \times \mathbf{s}(t) \cdot \mathbf{p}'(t) \]

6. If \( \mathbf{r}'(t) = 0 \) for all \( t \in (a, b) \), show there exists a constant vector, \( \mathbf{c} \) such that \( \mathbf{r}(t) = \mathbf{c} \) for all \( t \in (a, b) \).

   Do this by considering standard one variable calculus and on the components of \( \mathbf{r}(t) \).

7. If \( \mathbf{F}'(t) = \mathbf{f}(t) \) for all \( t \in (a, b) \) and \( \mathbf{F} \) is continuous on \([a, b]\), show \( \int_a^b \mathbf{f}(t) \, dt = \mathbf{F}(b) - \mathbf{F}(a) \).

   Do this by considering standard one variable calculus and on the components of \( \mathbf{r}(t) \).

8. Verify that if \( \mathbf{\Omega} \times \mathbf{u} = \mathbf{0} \) for all \( \mathbf{u} \), then \( \mathbf{\Omega} = \mathbf{0} \).

   Geometrically this says that if \( \mathbf{\Omega} \) is not equal to zero then it is parallel to every vector. Why does this make it obvious that \( \mathbf{\Omega} \) must equal zero?
Chapter 13

Newton’s Laws Of Motion*

I assume you have seen basic mechanics as found in introductory physics course. However, if you need a review, the following section is offered. Read it if you need to. Otherwise, skip it. Calculus was invented to solve problems in physics and engineering, not to do cute geometry. The material which follows on physics of motion on a space curve will make more sense to you if you know Newton’s laws.

**Definition 13.0.1** Let $\mathbf{r}(t)$ denote the position of an object. Then the acceleration of the object is defined to be $\mathbf{r}''(t)$.

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: 

$$F = m\mathbf{a} = m\mathbf{r}''(t)$$  \hfill (13.1)

where $\mathbf{a}$ 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 13.0.2** Let $\mathbf{r}(t)$ denote the position of an object of mass 2 kilogram at time $t$ and suppose the force acting on the object is given by $\mathbf{F}(t) = (t, 1 - t^2, 2e^{-t})$. Suppose $\mathbf{r}(0) = (1, 0, 1)$ meters, and $\mathbf{r}'(0) = (0, 1, 1)$ meters/sec. Find $\mathbf{r}(t)$.

---

1 Isaac Newton 1642-1727 is often credited with inventing calculus although this is not correct since most of the ideas were in existence earlier. However, he made major contributions to the subject partly in order to study physics and astronomy. He formulated the laws of gravity, made major contributions to optics, and stated the fundamental laws of mechanics listed here. He invented a version of the binomial theorem when he was only 23 years old and built a reflecting telescope. He showed that Kepler’s laws for the motion of the planets came from calculus and his laws of gravitation. In 1686 he published an important book, *Principia*, in which many of his ideas are found. Newton was also very interested in theology and had strong views on the nature of God which were based on his study of the Bible and early Christian writings. He finished his life as Master of the Mint.
By Newton’s second law, \(2r''(t) = F(t) = (t, 1 - t^2, 2e^{-t})\) and so
\[r''(t) = \left(\frac{t^2}{4} - \frac{t^3}{2}, -\frac{t^2}{2}, 2e^{-t}\right).
\]
Therefore the velocity is given by
\[r'(t) = \left(\frac{t^2}{4} - \frac{t^3}{2}, -\frac{t^2}{2}, 2e^{-t}\right) + c\]
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.
\[r'(t) = \left(\frac{t^2}{4} - \frac{t^3}{2} + 1, -e^{-t} + 2\right)\]
Now from this, the displacement must equal
\[r(t) = \left(\frac{t^3}{12}, \frac{t^2}{2} - \frac{t^4}{12}, t, e^{-t} + 2t\right) + (C_1, C_2, C_3)\]
where the constant vector, \((C_1, C_2, C_3)\) must be determined from the initial condition for the displacement. Thus
\[r(0) = (0, 0, 1) = (0, 0, 1) + (C_1, C_2, C_3)\]
which means \(C_1 = 1, C_2 = 0,\) and \(C_3 = 0\). Therefore, the displacement has also been found.
\[r(t) = \left(\frac{t^3}{12} + 1, \frac{t^2}{2} - \frac{t^4}{12} + t, e^{-t} + 2t\right) \text{ meters.}\]
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 13.0.3** 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.
\[m x''(t) = -mg, \quad r(0) = (0, 0), r'(0) = 400 \cos (\pi/6) \mathbf{i} + 400 \sin (\pi/6) \mathbf{j}.
\]
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 \( t \approx 40.8163265306 \) seconds and so at this time,
\[
x = 200\sqrt{3}(40.8163265306) = 14139.1902659 \text{ meters}.
\]
The next example is more complicated because it also takes in to account air resistance. We do not live in a vacume.

**Example 13.0.4** 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 \(-.1 r'(t)\) 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.
\[
64 = m \times 32.
\]
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\times5280 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
\[
(-.1) r'(t) \text{ pounds}.
\]

Newtons second law yields the following initial value problem for \( r(t) = (r_1(t), r_2(t)) \).
\[
2(r''_1(t), r''_2(t)) = (-.1)(r'_1(t), r'_2(t)) + (0, -64), \quad (r_1(0), r_2(0)) = (0, 30000), \quad (r'_1(0), r'_2(0)) = (880, 0)
\]
Therefore,
\[ 2r_1''(t) + (.1) r_1'(t) = 0 \]
\[ 2r_2''(t) + (.1) r_2'(t) = -64 \]
\[ r_1(0) = 0 \]
\[ r_2(0) = 30000 \]
\[ r_1'(0) = 880 \]
\[ r_2'(0) = 0 \] (13.2)

To save on repetition solve
\[ m x'' + k x = c, \quad x(0) = u, \quad x'(0) = v. \]

Divide the differential equation by \( m \) and get
\[ x'' + \left(\frac{k}{m}\right) x' = \frac{c}{m}. \]

Now multiply both sides by \( e^{\left(\frac{k}{m}\right)t} \). You should check this gives
\[ \frac{d}{dt} \left( e^{\left(\frac{k}{m}\right)t} x' \right) = \left(\frac{c}{m}\right) e^{\left(\frac{k}{m}\right)t} \]

Therefore,
\[ e^{\left(\frac{k}{m}\right)t} x' = \frac{1}{\left(\frac{k}{m}\right)} c + C \]

and using the initial condition, \( v = c/k + C \) and so
\[ x'(t) = \left(\frac{c}{k}\right) + \left(v - \left(\frac{c}{k}\right)\right) e^{-\frac{k}{m}t} \]

Now this implies
\[ x(t) = \left(\frac{c}{k}\right) t - \frac{1}{k} m e^{-\frac{k}{m}t} \left(v - \left(\frac{c}{k}\right)\right) + D \] (13.3)

where \( D \) is a constant to be determined from the initial conditions. Thus
\[ u = -\frac{m}{k} \left(v - \left(\frac{c}{k}\right)\right) + D \]

and so
\[ x(t) = \left(\frac{c}{k}\right) t - \frac{1}{k} m e^{-\frac{k}{m}t} \left(v - \left(\frac{c}{k}\right)\right) + \left(u + \frac{m}{k} \left(v - \left(\frac{c}{k}\right)\right)\right) . \]

Now apply this to the system (13.2) to find
\[ r_1(t) = -\frac{1}{(.1)} 2 \left( \exp \left( -\frac{(.1)}{2} t \right) \right) (880) + \left(\frac{2}{(.1)} \right) (880) \]
\[ = -17600.0 \exp (-.5t) + 17600.0 \]

and
\[ r_2(t) = (-64/(.1)) t - \frac{1}{(.1)} 2 \left( \exp \left( -\frac{(.1)}{2} t \right) \right) \left(\frac{64}{(.1)} \right) + \left(30000 + \frac{2}{(.1)} \left(\frac{64}{(.1)} \right) \right) \]
\[ = -640.0 t - 12800.0 \exp (-.5t) + 42800.0 \]

This gives the coordinates of the position. What of the velocity? Using (13.3) in the same way to obtain the velocity,
\[ r_1'(t) = 880.0 \exp (-.5t) , \]
\[ r_2'(t) = -640.0 + 640.0 \exp (-.5t) . \] (13.4)
13.1. KINETIC ENERGY

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 \( 13.4 \) to determine the velocity. It hits ground when \( r_2 (t) = 0 \). Thus it suffices to solve the equation,

\[
0 = -640.0t - 12800.0 \exp (-0.05t) + 42800.0.
\]

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 \( t = 66.14 \),

\[
-640.0 (66.14) - 12800.0 \exp (-0.05 (66.14)) + 42800.0 = 1.588 \text{ feet}.
\]

This is close enough to hitting the ground and so plugging in this value for \( t \) yields the approximate velocity,

\[
(880.0 \exp (-0.05 (66.14)), -640.0 + 640.0 \exp (-0.05 (66.14))) = (32.23, -616.56).
\]

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.

13.1 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 \( \mathbf{v} \) be the velocity,

\[
\mathbf{F} (t) = m\mathbf{v}' (t).
\]

Now in a small increment of time, \((t, t + dt)\), the work done on the object would be approximately equal to

\[
dW = \mathbf{F} (t) \cdot \mathbf{v} (t) \, dt. \tag{13.5}
\]

If no work has been done at time \( t = 0 \), then \( 13.5 \) implies

\[
\frac{dW}{dt} = \mathbf{F} \cdot \mathbf{v}, \quad W (0) = 0.
\]

Hence,

\[
\frac{dW}{dt} = m\mathbf{v}' (t) \cdot \mathbf{v} (t) = \frac{m}{2} \frac{d}{dt} |\mathbf{v} (t)|^2.
\]

Therefore, the total work done up to time \( t \) would be

\[
W (t) = \frac{m}{2} |\mathbf{v} (t)|^2 - \frac{m}{2} |\mathbf{v}_0|^2 \text{ where } |\mathbf{v}_0| \text{ denotes the initial speed of the object}. \]

This difference represents the change in the kinetic energy.
13.2 Impulse And Momentum*

Impulse

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 13.2.1 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 13.2.2 Let \( F \) be a force acting on an object of mass \( m \). Then the impulse equals the change in momentum. More precisely,

\[
\int_{a}^{b} F(t) \, dt = mv(b) - mv(a).
\]

Proof: This is really just the fundamental theorem of calculus and Newton’s second law applied to the components of \( F \).

\[
\int_{a}^{b} F(t) \, dt = \int_{a}^{b} m \frac{dv}{dt} \, dt = mv(b) - mv(a) \tag{13.6}
\]

13.3 Conservation Of Momentum*

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 \( v_{A} \) and \( v_{B} \) it follows that

\[
m_{A}v_{A}(b) - m_{A}v_{A}(a) = \int_{a}^{b} \text{(Force of } B \text{ on } A) \, dt
\]

and

\[
m_{B}v_{B}(b) - m_{B}v_{B}(a) = \int_{a}^{b} \text{(Force of } A \text{ on } B) \, dt
\]

\[
= -\int_{a}^{b} \text{(Force of } B \text{ on } A) \, dt
\]

\[
= -(m_{A}v_{A}(b) - m_{A}v_{A}(a))
\]

and this shows

\[
m_{B}v_{B}(b) + m_{A}v_{A}(a) = m_{B}v_{B}(a) + m_{A}v_{A}(a).
\]

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. This law is why rockets work. Think about it.
13.4 Exercises With Answers

1. Show the solution to $v' + rv = c$ with the initial condition, $v(0) = v_0$ is $v(t) = \left(v_0 - \frac{c}{r}\right)e^{-rt} + \frac{c}{r}$. If $v$ is velocity and $r = k/m$ where $k$ is a constant for air resistance and $m$ is the mass, and $c = f/m$, argue from Newton’s second law that this is the equation for finding the velocity, $v$ of an object acted on by air resistance proportional to the velocity and a constant force, $f$, possibly from gravity. Does there exist a terminal velocity? What is it?

Multiply both sides of the differential equation by $e^{rt}$. Then the left side becomes

$$\frac{d}{dt}(e^{rt}v) = e^{rt}c.$$ 

Now integrate both sides. This gives

$$e^{rt}v(t) = C + \frac{c}{r}e^{rt}.$$ 

You finish the rest.

2. Suppose an object having mass equal to 5 kilograms experiences a time dependent force, $F(t) = e^{-t}i + \cos(t)j + t^2k$ 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 $v = i + j - k$ meters per sec. Find the position of the object as a function of $t$.

This is done by using Newton’s law. Thus $5\frac{d^2r}{dt^2} = e^{-t}i + \cos(t)j + t^2k$ and so

$$5\frac{dr}{dt} = -e^{-t}i + \sin(t)j + \left(t^3/3\right)k + C.$$ 

Find the constant, $C$ by using the given initial velocity. Next do another integration obtaining another constant vector which will be determined by using the given initial position of the object.

3. 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 $dW = \mathbf{F}(t) \cdot \mathbf{v}(t) dt$?

Remember $|\mathbf{v}|^2 = \mathbf{v} \cdot \mathbf{v}$. Now use the product rule.

4. 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 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.

$$0 = \mathbf{F} \cdot \mathbf{v} = m\mathbf{v}' \cdot \mathbf{v}.$$ 

Explain why this is $\frac{d}{dt} \left( m \frac{1}{2} |\mathbf{v}|^2 \right)$, the derivative of the kinetic energy.
Chapter 14

Physics Of Curvilinear Motion

14.0.1 The Acceleration In Terms Of The Unit Tangent And Normal

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 \( \mathbf{R}(t) \) the position vector of the point on the curve which occurs at time \( t \). Assume that \( \mathbf{R}', \mathbf{R}'' \) exist and is continuous. Thus \( \mathbf{R}' = \mathbf{v} \), the velocity and \( \mathbf{R}'' = \mathbf{a} \) is the acceleration.

Thus

\[
\mathbf{v} = |\mathbf{v}| \mathbf{T}
\]

where \( \mathbf{T} \) is the unit tangent vector. The acceleration is \( \mathbf{v}' = \mathbf{a} \). The idea is to write \( \mathbf{a} \) as the sum of two vectors, one which is in the direction of \( \mathbf{T} \) and the other which is perpendicular to \( \mathbf{T} \). That is,

\[
\mathbf{a} = a_T \mathbf{T} + a_N \mathbf{N}, \ |\mathbf{T}| = |\mathbf{N}| = 1, \ \mathbf{T} \cdot \mathbf{N} = 0, \ a_N \geq 0
\]

so what is \( a_T \) and \( a_N \)? First take the dot product of both sides with \( \mathbf{T} \). Then

\[
\mathbf{a} \cdot \mathbf{T} = a_T \mathbf{T} \cdot \mathbf{T} = a_T.
\]

Then

\[
\mathbf{a} = (\mathbf{a} \cdot \mathbf{T}) \mathbf{T} + a_N \mathbf{N}
\]

\[
\mathbf{a} - (\mathbf{a} \cdot \mathbf{T}) \mathbf{T} = a_N \mathbf{N}
\]

\[
|a_N| = a_N = |\mathbf{a} - (\mathbf{a} \cdot \mathbf{T}) \mathbf{T}|
\]
and so
\[ \mathbf{N} = \frac{\mathbf{a} - (\mathbf{a} \cdot \mathbf{T}) \mathbf{T}}{|\mathbf{a} - (\mathbf{a} \cdot \mathbf{T}) \mathbf{T}|} \]

This is called the principal normal vector.

\[ \mathbf{a} = (\mathbf{a} \cdot \mathbf{T}) \mathbf{T} + |\mathbf{a} - (\mathbf{a} \cdot \mathbf{T}) \mathbf{T}| \frac{\mathbf{a} - (\mathbf{a} \cdot \mathbf{T}) \mathbf{T}}{|\mathbf{a} - (\mathbf{a} \cdot \mathbf{T}) \mathbf{T}|} \]

= \( (\mathbf{a} \cdot \mathbf{T}) \mathbf{T} + |\mathbf{a} - (\mathbf{a} \cdot \mathbf{T}) \mathbf{T}| \mathbf{N} \)

Now note that \( a_T \) can be either positive or negative but \( a_N \) is always nonnegative. Also,

\[ |\mathbf{a}|^2 = (a_T \mathbf{T} + a_N \mathbf{N}) \cdot (a_T \mathbf{T} + a_N \mathbf{N}) \]

= \( a_T^2 \mathbf{T} \cdot \mathbf{T} + 2a_N a_T \mathbf{T} \cdot \mathbf{N} + a_N^2 \mathbf{N} \cdot \mathbf{N} \)

= \( a_T^2 + a_N^2 \)

Therefore, if you have found \( a_T \), you can determine \( a_N \) as

\[ a_N = \sqrt{|\mathbf{a}|^2 - a_T^2} \]

**Example 14.0.1** Let \( \mathbf{R}(t) = (t, t^2 + \frac{t^3}{3}) \). Find the acceleration in the form \( \mathbf{a} = a_T \mathbf{T} + a_N \mathbf{N} \) where \( a_N \geq 0 \) and \( \mathbf{T} \) is the unit tangent vector while \( \mathbf{N} \) is the principal normal vector.

First find \( \mathbf{T} \). This equals

\[ \mathbf{T} = \frac{(1, t, t^2)}{\sqrt{1 + t^2 + t^4}} \]

Next,

\[ \mathbf{a} = (0, 1, 2t) \]

Then

\[ a_T = (\mathbf{a} \cdot \mathbf{T}) = \frac{t + 2t^3}{\sqrt{1 + t^2 + t^4}} \]

So what is \( a_N \)?

\[ a_N = \sqrt{1 + 4t^2 - \left( \frac{t + 2t^3}{\sqrt{1 + t^2 + t^4}} \right)^2} \]

= \( \sqrt{t^4 + 4t^2 + 1} \)

Then

\[ \mathbf{a} = \frac{t + 2t^3}{\sqrt{1 + t^2 + t^4}} \mathbf{T} + \sqrt{t^4 + 4t^2 + 1} \frac{N}{t^2 + t^2 + 1} \]

It is customary to massage \( a_N \) some more to write it in terms of a geometrical property of the curve which is known as curvature, denoted as \( \kappa \).

**Lemma 14.0.2** Define \( \mathbf{T}(t) \equiv \mathbf{R}'(t) / |\mathbf{R}'(t)| \). Then \( |\mathbf{T}(t)| = 1 \) and if \( \mathbf{T}'(t) \neq 0 \), then there exists a unit vector, \( \mathbf{N}(t) \) perpendicular to \( \mathbf{T}(t) \) and a scalar valued function, \( \kappa(t) \), with \( \mathbf{T}'(t) = \kappa(t) |\mathbf{v}| \mathbf{N}(t) \).
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.

Definition 14.0.3 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 \(\rho = 1/\kappa\). 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.

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 14.0.4 For \(\mathbf{R}(t)\) the position vector of a space curve, the acceleration is given by the formula
\[
a = \frac{d|v|}{dt} T + \kappa |v|^2 N \quad \equiv \quad a_T T + a_N N.
\]
Furthermore, \(a_T^2 + a_N^2 = |a|^2\).

Proof: The last claim was considered above.
\[
a = \frac{dv}{dt} = \frac{d}{dt} (\mathbf{R}') = \frac{d}{dt} (|v| T)
= \frac{d|v|}{dt} T + |v| T'
= \frac{d|v|}{dt} T + |v|^2 \kappa N.
\]
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 parametrization of the curve. If the curve is given by two different vector valued functions, \(\mathbf{R}(t)\) and \(\mathbf{R}(\tau)\), then from the formula above for the curvature,
\[
\kappa(t) = \frac{|T'(t)|}{|v(t)|} = \frac{\left|\frac{dT}{d\tau} \frac{d\tau}{dt}\right|}{\left|\frac{d\mathbf{R}}{d\tau} \frac{d\tau}{dt}\right|} = \left|\frac{dT}{d\tau}\right| \equiv \kappa(\tau).
\]
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

\[\text{To osculate means to kiss. Thus this plane could be called the kissing plane. However, that does not sound formal enough so we call it the osculating plane.} \]
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 parametrization of the object which is as described is

$$R(t) = \left( r \cos \left( \frac{v}{r} t \right), r \sin \left( \frac{v}{r} t \right) \right).$$

Therefore, $T = (- \sin \left( \frac{v}{r} t \right), \cos \left( \frac{v}{r} t \right))$ and $T' = \left( - \frac{v}{r} \cos \left( \frac{v}{r} t \right), - \frac{v}{r} \sin \left( \frac{v}{r} t \right) \right)$. Thus,

$$\kappa = \frac{|T'(t)|}{v} = \frac{1}{r}.$$

I hope it is not surprising that the curvature of a circle of radius $r$ is $1/r$. It follows

$$a = \frac{dv}{dt} T + v^2 \kappa N = \frac{v^2}{r} N.$$

The vector, $N$ 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 $v$, the velocity. Then

$$a \times v = \frac{d|v|}{dt} T \times v + |v|^2 \kappa N \times v = \frac{d|v|}{dt} T \times v + |v|^3 \kappa N \times T$$

Now $T$ and $v$ have the same direction so the first term on the right equals zero. Taking the magnitude of both sides, and using the fact that $N$ and $T$ are two perpendicular unit vectors,

$$|a \times v| = |v|^3 \kappa$$

and so

$$\kappa = \frac{|a \times v|}{|v|^3}. \quad (14.2)$$

**Example 14.0.5** Let $R(t) = (\cos(t), t, t^2)$ for $t \in [0, 3]$. Find the speed, velocity, curvature, and write the acceleration in terms of normal and tangential components.

First of all $v(t) = (-\sin t, 1, 2t)$ and so the speed is given by

$$|v| = \sqrt{\sin^2(t) + 1 + 4t^2}.$$

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.

$$a(t) = R''(t) = (-\cos t, 0, 2).$$

Then

$$\kappa = \frac{|(-\cos t, 0, 2) \times (-\sin t, 1, 2t)|}{\left( \sqrt{\sin^2(t) + 1 + 4t^2} \right)^3}$$

$$= \frac{\sqrt{4 + (-2 \sin(t) + 2(\cos(t)) t)^2 + \cos^2(t)}}{\left( \sqrt{\sin^2(t) + 1 + 4t^2} \right)^3}$$
Then
\[ a_N = \kappa |\mathbf{v}|^2 = \sqrt{4 + (-2 \sin (t) + 2 \cos (t) t)^2 + \cos^2 (t)} \left( \frac{\sin^2 (t) + 1 + 4t^2}{\sqrt{\sin^2 (t) + 1 + 4t^2}} \right)^3 \]
\[ = \sqrt{4 + (-2 \sin (t) + 2 \cos (t) t)^2 + \cos^2 (t)} \]
\[ \sqrt{\sin^2 (t) + 1 + 4t^2} \].

You can observe the formula \( a_N^2 + a_T^2 = |a|^2 \) holds. Indeed \( a_N^2 + a_T^2 = \)
\[ \left( \frac{\sqrt{4 + (-2 \sin (t) + 2 \cos (t) t)^2 + \cos^2 (t)}}{\sin^2 (t) + 1 + 4t^2} \right)^2 + \left( \frac{\sin (t) \cos (t) + 4t}{\sqrt{2 + 4t^2 - \cos^2 t}} \right)^2 \]
\[ = \frac{4 + (-2 \sin t + 2 \cos (t) t)^2 + \cos^2 t}{\sin^2 t + 1 + 4t^2} + \frac{(\sin t \cos t + 4t)^2}{2 + 4t^2 - \cos^2 t} = \cos^2 t + 4 = |a|^2. \]

Some Examples

To illustrate the use of these simple observations, consider the example worked above which was fairly messy. I will make it easier by selecting a value of \( t \).

**Example 14.0.6** 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 when \( t = 0 \). Also find \( \mathbf{N} \) at the point where \( t = 0 \).

First I need to find the velocity and acceleration. Thus
\[ \mathbf{v} = (- \sin t, 1, 2t), \quad \mathbf{a} = (- \cos t, 0, 2) \]
and consequently,
\[ \mathbf{T} = \frac{(- \sin t, 1, 2t)}{\sqrt{\sin^2 (t) + 1 + 4t^2}}. \]

When \( t = 0 \), this reduces to
\[ \mathbf{v}(0) = (0, 1, 0), \quad \mathbf{a} = (-1, 0, 2), \quad |\mathbf{v}(0)| = 1, \quad \mathbf{T} = (0, 1, 0), \]
and consequently,
\[ \mathbf{T} = (0, 1, 0). \]

Then the tangential component of acceleration when \( t = 0 \) is
\[ a_T = (-1, 0, 2) \cdot (0, 1, 0) = 0 \]
Now \( |\mathbf{a}|^2 = 5 \) and so \( a_N = \sqrt{5} \) because \( a_T^2 + a_N^2 = |\mathbf{a}|^2 \). Thus \( \sqrt{5} = \kappa |\mathbf{v}(0)|^2 = \kappa \cdot 1 = \kappa \).

Next let's find \( \mathbf{N} \). From \( \mathbf{a} = a_T \mathbf{T} + a_N \mathbf{N} \) it follows
\[ (-1, 0, 2) = 0 \cdot \mathbf{T} + \sqrt{5} \mathbf{N} \]
and so
\[ \mathbf{N} = \frac{1}{\sqrt{5}} (-1, 0, 2). \]

This was pretty easy.
Example 14.0.7 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 parametrization 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,

\[
a \times v = (0, f''(t), 0) \times (1, f'(t), 0) = (0, 0, -f''(t)).
\]

Therefore, the curvature is given by

\[
\frac{|a \times v|}{|v|^3} = \frac{|f''(t)|}{\left(1 + f'(t)^2\right)^{3/2}}.
\]

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 parametrization. General systems of nonlinear equations cannot be solved using algebra.

Example 14.0.8 Find a parametrization 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

\[
z = 2x^2 - x + 3, \quad y = -4x^2 + 3x - 5.
\]

Therefore, letting \( t = x \), the parametrization is \((x, y, z) = (t, -4t^2 - 5 + 3t, -t + 3 + 2t^2)\).

Example 14.0.9 Find a parametrization 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.

Example 14.0.10 What does it mean if \( r(t), t \in [a, b] \) is a space curve and

\[
|r'(t)| = 1
\]

for all \( t \)? For example if

\[
r(t) = \left(\frac{\sqrt{2}}{2} \cos(t), \frac{\sqrt{2}}{2} \sin(t), \frac{\sqrt{2}}{2} t\right), \quad t \in [1, 2]
\]

a short computation shows

\[
|r'(t)|^2 = \frac{1}{2} + \frac{1}{2} = 1
\]

so this is such an example.
Letting \( s \) denote the arc length of the arc traced out by \( r(t) \) between the points \( r(c) \) and \( r(d) \). This equals
\[
s = \int_c^d |r'(t)| \, dt = (d - c).
\]
Thus the arc length of a segment of the curve is equal to the length of the time interval corresponding to this segment of the curve. When this happens, we say the curve is \textbf{parametrized with respect to arc length}. When one considers the full theory of space curves, the concepts are presented in terms of the parametrization with respect to arc length. This is done because it becomes more obvious that certain properties are intrinsic properties of the curve itself and are not dependent on the parametrization used. See below if you are interested in the whole story.

Given a space curve having parametrization \( r(t), t \in [a, b] \), you can always (theoretically) change the parameter to obtain the same space curve parametrized with respect to arc length. Let
\[
L = \int_a^b |r'(t)| \, dt
\]
the length of the whole curve. Then define a new parameter,
\[
s \equiv \phi(t) = \int_a^t |r'(\tau)| \, d\tau
\]
Then assuming this is a smooth curve, it follows \( \phi'(t) = |r'(t)| > 0 \) and so \( \phi(t) \) is a one to one function, \( \phi : [a, b] \to [0, L] \). Then \( s \to \phi^{-1}(s) \) is also a smooth function and you can consider
\[
R(s) \equiv r(\phi^{-1}(s))
\]
It traces out the same points in the same order but
\[
|R'(s)| = \left| r' \left( \phi^{-1}(s) \right) \left( \phi^{-1} \right)'(s) \right|
\]
\[
= \left| r' \left( \phi^{-1}(s) \right) \frac{1}{\phi'(\phi^{-1}(s))} \right| = \left| \frac{r' \left( \phi^{-1}(s) \right)}{|r'(\phi^{-1}(s))|} \right| = 1
\]
so this is a parametrization of the curve with respect to arc length. More of this is discussed below in the section on the geometric theory of space curves.

\section{14.0.2 The Circle Of Curvature*}

In addition to the osculating plane, you can consider something called the circle of curvature. The idea is that near a point on the space curve, the space curve is like a circle. This circle has radius equal to \( 1/\kappa \), the radius of curvature, lies in the osculating plane, and its center is located by moving a distance of \( 1/\kappa \) (radius of curvature) along the line determined by the point on the curve and the principle normal in the direction of the principle normal. It is an attempt to find the circle which best resembles the curve locally.
Here is an example to illustrate this fussy concept.

**Example 14.0.11** Consider the curve having a parametrization, \((\cos(t), \sin(t), e^t)\). Find the circle of curvature at the point where \(t = \pi/4\).

First find the curvature and the two vectors, \(T, N\). The vector, 
\[
T(t) = \left( -\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, e^{\pi/4} \right) \frac{1}{\sqrt{1 + e^{2t}}}.
\]

To find \(N\) next appears to be painful. Therefore, I will first find the acceleration and then use the formula for acceleration to dredge up \(N\).

\[
a(t) = (-\cos t, -\sin t, e^t)
\]

Thus the curvature is easy to find.

\[
\kappa = \frac{\left| (-\cos t, -\sin t, e^t) \times (-\sin t, \cos t, e^t) \right|}{(\sqrt{1 + e^{2t}})^3}
\]

\[
= \frac{\left| (-\sin t) e^t - e^t \cos t, - (\sin t) e^t + e^t \cos t, -\cos^2 t - \sin^2 t \right|}{(\sqrt{1 + e^{2t}})^3}
\]

\[
= \frac{(2e^{2t} + 1)^{1/2}}{(\sqrt{1 + e^{2t}})^3}
\]

It follows that at the point of interest,

\[
\left( -\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}, e^{\pi/4} \right) = \left( -\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, e^{\pi/4} \right) + \frac{(2e^{\pi/2} + 1)^{1/2}}{(\sqrt{1 + e^{\pi/2}})^3} \left( 1 + e^{\pi/2} \right) N
\]

From this, you can find the components of \(N\) without too much trouble.

\[
-\frac{\sqrt{2}}{2} = -\frac{\sqrt{2}}{2} + \frac{(2e^{\pi/2} + 1)^{1/2}}{(\sqrt{1 + e^{\pi/2}})^3} \left( 1 + e^{\pi/2} \right) N
\]
and so $N_1 = 0$. Next,

$$-\sqrt{2} = \frac{\sqrt{2}}{2} + \frac{(2e^{\pi/2} + 1)^{1/2}}{\sqrt{1 + e^{\pi/2}}} \left(1 + e^{\pi/2}\right) N_2$$

and so

$$N_2 = -\sqrt{2} \frac{\sqrt{1 + e^{\pi/2}}}{\sqrt{(2e^{\pi/2} + 1)}}$$

Finally,

$$e^{\pi/4} = e^{\pi/4} + \frac{(2e^{\pi/2} + 1)^{1/2}}{\sqrt{1 + e^{\pi/2}}} \left(1 + e^{\pi/2}\right) N_3$$

and so $N_3 = 0$ also.

From this you can find the location of the center of curvature. It is at the point

$$\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, e^{\pi/4}\right) + \left(\frac{2e^{\pi/2} + 1}{\sqrt{1 + e^{\pi/2}}} \right)^{-1} \left(0, -\sqrt{2} \frac{\sqrt{1 + e^{\pi/2}}}{\sqrt{(2e^{\pi/2} + 1)}}, 0\right).$$

Simplifying this yields

$$\left(\frac{1}{2} \sqrt{2}, -\frac{1}{2} \sqrt{22e^{\pi/2} + 1 + 2e^{\pi/2}}, e^{\pi/4}\right)$$

for the center of curvature. The circle of curvature is the circle in the osculating plane which has the above point as the center and radius equal to

$$\frac{1}{\left(\frac{2e^{\pi/2} + 1}{\sqrt{1 + e^{\pi/2}}} \right)^{-1}} = \frac{1}{\sqrt{\left(2e^{\pi/2} + 1\right)}} \left(\sqrt{1 + e^{\pi/2}}\right)^3.$$

If you try the same problem for $(\cos(t), \sin(t), t)$, you may find the computations much simpler.

One can of course go on and on fussing about geometrical dodads of this sort. The evolute is the locus of centers of curvatures. Imagine finding such a center of curvature as above for each $t$ and considering the resulting curve. This is the evolute. Then if you really like to do this sort of thing, you could think about the evolute of the evolute. There are sure to be some wonderful conclusions hidden in this procedure.

The significant geometrical concepts are discussed in Section 14.1. These lead to the Serrat Frenet formulas which cause some people who like this sort of thing to wax ecstatic over their virtues. These formulas are indeed interesting, unlike the fussy stuff above about the circle of curvature. However, it is not required reading so skip it if you are not interested. It is a system of differential equations which completely describes the geometry of the space curve.

### 14.1 Geometry Of Space Curves

If you are interested in more on space curves, you should read this section. Otherwise, proceed to the exercises. Denote by $\mathbf{R}(s)$ the function which takes $s$ to a point on this
curve where \( s \) is arc length. Thus \( \mathbf{R}(s) \) equals the point on the curve which occurs when you have traveled a distance of \( s \) along the curve from one end. This is known as the parametrization of the curve in terms of arc length. Note also that it incorporates an orientation on the curve because there are exactly two ends you could begin measuring length from. In this section, assume anything about smoothness and continuity to make the following manipulations valid. In particular, assume that \( \mathbf{R'} \) exists and is continuous.

**Lemma 14.1.1** Define \( \mathbf{T}(s) \equiv \mathbf{R'}(s) \). Then \( |\mathbf{T}(s)| = 1 \) and if \( \mathbf{T}'(s) \neq 0 \), then there exists a unit vector, \( \mathbf{N}(s) \) perpendicular to \( \mathbf{T}(s) \) and a scalar valued function, \( \kappa(s) \) with \( \mathbf{T}'(s) = \kappa(s) \mathbf{N}(s) \).

**Proof:** First, \( s = \int_{a}^{s} |\mathbf{R}'(r)| \, dr \) because of the definition of arc length. Therefore, from the fundamental theorem of calculus, \( 1 = |\mathbf{R}'(s)| = |\mathbf{T}(s)| \). Therefore, \( \mathbf{T} \cdot \mathbf{T} = 1 \) and so upon differentiating this on both sides, yields \( \mathbf{T}' \cdot \mathbf{T} + \mathbf{T} \cdot \mathbf{T}' = 0 \) which shows \( \mathbf{T} \cdot \mathbf{T}' = 0 \). Therefore, the vector, \( \mathbf{T}' \) is perpendicular to the vector, \( \mathbf{T} \). In case \( \mathbf{T}'(s) \neq 0 \), let \( \mathbf{N}(s) = \frac{\mathbf{T}'(s)}{|\mathbf{T}'(s)|} \) and so \( \mathbf{T}'(s) = |\mathbf{T}'(s)| \mathbf{N}(s) \), showing the scalar valued function is \( \kappa(s) = |\mathbf{T}'(s)| \). This proves the lemma.

The radius of curvature is defined as \( \rho = \frac{1}{\kappa} \). Thus at points where there is a lot of curvature, the radius of curvature is small and at points where the curvature is small, the radius of curvature is large. The plane determined by the two vectors, \( \mathbf{T} \) and \( \mathbf{N} \) is called the osculating plane. It identifies a particular plane which is in a sense tangent to this space curve. In the case where \( |\mathbf{T}'(s)| = 0 \) near the point of interest, \( \mathbf{T}(s) \) 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 \( |\mathbf{T}'(s)| = 0 \), \( \kappa(s) = 0 \) and the radius of curvature would be considered infinite.

**Definition 14.1.2** The vector, \( \mathbf{T}(s) \) is called the unit tangent vector and the vector, \( \mathbf{N}(s) \) is called the principal normal. The function, \( \kappa(s) \) in the above lemma is called the curvature. When \( \mathbf{T}'(s) \neq 0 \) so the principal normal is defined, the vector, \( \mathbf{B}(s) \equiv \mathbf{T}(s) \times \mathbf{N}(s) \) is called the binormal.

The binormal is normal to the osculating plane and \( \mathbf{B}' \) tells how fast this vector changes. Thus it measures the rate at which the curve twists.

**Lemma 14.1.3** Let \( \mathbf{R}(s) \) be a parametrization of a space curve with respect to arc length and let the vectors, \( \mathbf{T}, \mathbf{N}, \) and \( \mathbf{B} \) be as defined above. Then \( \mathbf{B}' = \mathbf{T} \times \mathbf{N}' \) and there exists a scalar function, \( \tau(s) \) such that \( \mathbf{B}' = \tau \mathbf{N} \).

**Proof:** From the definition of \( \mathbf{B} = \mathbf{T} \times \mathbf{N} \), and you can differentiate both sides and get \( \mathbf{B}' = \mathbf{T}' \times \mathbf{N} + \mathbf{T} \times \mathbf{N}' \). Now recall that \( \mathbf{T}' \) is a multiple called curvature multiplied by \( \mathbf{N} \) so the vectors, \( \mathbf{T}' \) and \( \mathbf{N} \) have the same direction and \( \mathbf{B}' = \mathbf{T} \times \mathbf{N}' \). Therefore, \( \mathbf{B}' \) is either zero or is perpendicular to \( \mathbf{T} \). But also, from the definition of \( \mathbf{B}, \mathbf{B} \) is a unit vector and so \( \mathbf{B}(s) \cdot \mathbf{B}(s) = 0 \). Differentiating this, \( \mathbf{B}'(s) \cdot \mathbf{B}(s) + \mathbf{B}(s) \cdot \mathbf{B}'(s) = 0 \) showing that \( \mathbf{B}' \) is perpendicular to \( \mathbf{B} \) also. Therefore, \( \mathbf{B}' \) is a vector which is perpendicular to both vectors, \( \mathbf{T} \) and \( \mathbf{B} \) and since this is in three dimensions, \( \mathbf{B}' \) must be some scalar multiple of \( \mathbf{N} \) and it is this multiple called \( \tau \). Thus \( \mathbf{B}' = \tau \mathbf{N} \) as claimed.

Lets go over this last claim a little more. The following situation is obtained. There are two vectors, \( \mathbf{T} \) and \( \mathbf{B} \) which are perpendicular to each other and both \( \mathbf{B}' \) and \( \mathbf{N} \) are perpendicular to these two vectors, hence perpendicular to the plane determined by them. Therefore, \( \mathbf{B}' \) must be a multiple of \( \mathbf{N} \). Take a piece of paper, draw two unit vectors on it
which are perpendicular. Then you can see that any two vectors which are perpendicular
to this plane must be multiples of each other.

The scalar function, \( \tau \) is called the torsion. In case \( T' = 0 \), none of this is defined
because in this case there is not a well defined osculating plane. The conclusion of the
following theorem is called the Serret Frenet formulas.

**Theorem 14.1.4 (Serret Frenet)** Let \( R(s) \) be the parametrization with respect to
arc length of a space curve and \( T(s) = T'(s) \) is the unit tangent vector. Suppose \( |T'(s)| \neq 0 \)
so the principal normal, \( N(s) = \frac{T'(s)}{|T'(s)|} \) is defined. The binormal is the vector \( B = T \times N \)
so \( T, N, B \) forms a right handed system of unit vectors each of which is perpendicular
to every other. Then the following system of differential equations holds in \( \mathbb{R}^3 \).

\[
B' = \tau N, \quad T' = \kappa N, \quad N' = -\kappa T - \tau B
\]

where \( \kappa \) is the curvature and is nonnegative and \( \tau \) is the torsion.

**Proof:** \( \kappa \geq 0 \) because \( \kappa = |T'(s)| \). The first two equations are already established. To
get the third, note that \( B \times T = N \) which follows because \( T, N, B \) is given to form a right
handed system of unit vectors each perpendicular to the others. (Use your right hand.)
Now take the derivative of this expression. thus

\[
N' = B' \times T + B \times T' = \tau N \times T + \kappa B \times N.
\]

Now recall again that \( T, N, B \) is a right hand system. Thus \( N \times T = -B \) and \( B \times N = -T \).
This establishes the Frenet Serret formulas.

This is an important example of a system of differential equations in \( \mathbb{R}^3 \). It is a remarkable
result because it says that from knowledge of the two scalar functions, \( \tau \) and \( \kappa \), and initial
values for \( B, T \), and \( N \) when \( s = 0 \) you can obtain the binormal, unit tangent, and principal
normal vectors. It is just the solution of an initial value problem of the sort discussed
earlier. Having done this, you can reconstruct the entire space curve starting at some point,
\( R_0 \) because \( R'(s) = T(s) \) and so \( R(s) = R_0 + \int_0^s T'(r) \, dr \).

The vectors, \( B, T \), and \( N \) are vectors which are functions of position on the space curve.
Often, especially in applications, you deal with a space curve which is parametrized by a
function of \( t \) where \( t \) is time. Thus a value of \( t \) would correspond to a point on this curve and
you could let \( B(t), T(t) \), and \( N(t) \) be the binormal, unit tangent, and principal normal at
this point of the curve. The following example is typical.

**Example 14.1.5** Given the circular helix, \( R(t) = (a \cos t) \, i + (a \sin t) \, j + (bt) \, k \), find the
arc length, \( s(t) \), the unit tangent vector, \( T(t) \), the principal normal, \( N(t) \), the binormal,
\( B(t) \), the curvature, \( \kappa(t) \), and the torsion, \( \tau(t) \). Here \( t \in [0, T] \).

The arc length is \( s(t) = \int_0^t \sqrt{a^2 + b^2} \, dr = (\sqrt{a^2 + b^2} \, t). \) Now the tangent vector
is obtained using the chain rule as

\[
T = \frac{dR}{ds} \, dt \, ds = \frac{1}{\sqrt{a^2 + b^2}} R'(t)
\]

\[
= \frac{1}{\sqrt{a^2 + b^2}} \left( (-a \sin t) \, i + (a \cos t) \, j + bk \right)
\]

The principal normal:

\[
\frac{dT}{ds} = \frac{dT}{dt} \, dt \, ds = \frac{1}{a^2 + b^2} \left( (-a \cos t) \, i + (-a \sin t) \, j + 0k \right)
\]
CHAPTER 14. PHYSICS OF CURVILINEAR MOTION

and so

\[ N = \frac{dT}{ds} \times \frac{dT}{ds} = -((\cos t) \mathbf{i} + (\sin t) \mathbf{j}) \]

The binormal:

\[
B = \frac{1}{\sqrt{a^2 + b^2}} \begin{vmatrix}
\mathbf{i} & \mathbf{j} & \mathbf{k} \\
-a \sin t & a \cos t & b \\
-\cos t & -\sin t & 0
\end{vmatrix}
\]

= \frac{1}{\sqrt{a^2 + b^2}} ((b \sin t) \mathbf{i} - b \cos t \mathbf{j} + a \mathbf{k})

Now the curvature, \( \kappa(t) = \left| \frac{dT}{ds} \right| = \sqrt{\left(\frac{a \cos t}{a^2 + b^2}\right)^2 + \left(\frac{a \sin t}{a^2 + b^2}\right)^2} = \frac{a}{\sqrt{a^2 + b^2}}. \) Note the curvature is constant in this example. The final task is to find the torsion. Recall that \( B' = \tau N \) where the derivative on \( B \) is taken with respect to arc length. Therefore, remembering that \( t \) is a function of \( s \),

\[
B'(s) = \frac{1}{\sqrt{a^2 + b^2}} ((b \cos t) \mathbf{i} + (b \sin t) \mathbf{j}) \frac{dt}{ds}
\]

= \frac{1}{\sqrt{a^2 + b^2}} ((b \cos t) \mathbf{i} + (b \sin t) \mathbf{j})

= \tau (-\cos t \mathbf{i} - \sin t \mathbf{j}) = \tau N

and it follows \(-b/ (a^2 + b^2) = \tau.\)

An important application of the usefulness of these ideas involves the decomposition of the acceleration in terms of these vectors of an object moving over a space curve.

**Corollary 14.1.6** Let \( R(t) \) be a space curve and denote by \( v(t) \) the velocity, \( v(t) = R'(t) \) and let \( v(t) \equiv |v(t)| \) denote the speed and let \( a(t) \) denote the acceleration. Then \( v = vT \) and \( a = \frac{dv}{dt} = \frac{dv}{dt} T + v^2 \kappa N. \)

**Proof:** \( T = \frac{dR}{ds} = \frac{dR}{dt} \frac{dt}{ds} = v \frac{dt}{ds}. \) Also, \( s = \int_0^t v(\tau) \, d\tau \) and so \( \frac{dt}{ds} = \frac{1}{v}. \) Therefore, \( T = v/v \) which implies \( v = vT \) as claimed.

Now the acceleration is just the derivative of the velocity and so by the Serrat Frenet formulas,

\[
a = \frac{dv}{dt} T + v \frac{dT}{dt} \]

= \frac{dv}{dt} T + \frac{dv}{ds}v = \frac{dv}{dt} T + v^2 \kappa N

Note how this decomposes the acceleration into a component tangent to the curve and one which is normal to it. Also note that from the above, \( v |T'| \frac{T'(t)}{|T'|} = v^2 \kappa N \) and so \( \frac{T'}{v} = \kappa \) and \( N = \frac{T'}{|T'|} \)

14.2 Independence Of Parametrization*
This section is for those who want to really understand what is going on. If you are content, do not read this section. It may upset you. However, if you do decide to read it, you might learn something so there is some benefit for the anguish you might endure in the attempt.

Recall that if \( p(t) : t \in [a, b] \) was a parametrization of a smooth curve, \( C \), the length of \( C \) is defined as

\[
\int_a^b |p'(t)| \, dt
\]

If some other parametrization were used to trace out \( C \), would the same answer be obtained? The answer is yes. This is indeed fortunate because the length of a curve should only depend on the curve itself, not on some parametrization. To answer this question in a satisfactory manner requires some hard calculus. To answer it even more satisfactorily, you need to consider some very advanced mathematics involving something called Hausdorff measure.

### 14.2.1 Hard Calculus

**Definition 14.2.1**

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. The limit of a sequence, if it exists, is unique.

**Theorem 14.2.2**

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 \( 0 < \varepsilon < |a_1 - a|/2 \) 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 14.2.3**

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 \( b_k = a_{n_k} \), \( \{b_k\} \) is called a subsequence of \( \{a_n\} \).

**Theorem 14.2.4**

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 14.2.5 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 $|x - x_n| < \frac{1}{n}$, 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 14.2.6 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 14.2.7 Every closed interval, $[a, b]$ is sequentially compact.

Proof: Let $\{x_n\} \subseteq [a, b] \equiv I_0$. Consider the two intervals $[a, \frac{a+b}{2}]$ and $[\frac{a+b}{2}, b]$ 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 $x_{n_3} \in I_3$, etc. (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.

Theorem 14.2.8 If $f : [a, b] \to \mathbb{R}$ is continuous and $f(b) > 0, f(0) < 0$, then there exists $t \in (a, b)$ such that $f(t) = 0$.

Proof: Let $S \equiv \{t \in [a, b] : f(s) \leq 0 \text{ for all } s \in [a, t]\}$. Then $[a, a] \subseteq S$. Let $t$ be the least upper bound of $S$ which exists by the assumption that $\mathbb{R}$ is complete. Then $t \in (a, b)$ because, by continuity, $f(s) < 0$ for all $s$ close enough to $a$. Similarly $t < b$ because of similar reasoning. If $f(t) \neq 0$, then by continuity, there exists $\delta > 0$ such that if $s \in (t - \delta, t + \delta) \subseteq (a, b)$ it follows that $f(s)$ is also nonzero and has the same sign as $f(t)$. Hence $t$ is either not an upper bound if $f(t) < 0$ or it is not a least upper bound if $f(t) > 0$. This proves the theorem.

This theorem is called the intermediate value theorem. It is what makes it possible to prove the following lemma.
Lemma 14.2.9 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

$$\left(\phi(y_1) - \phi(x_1)\right)(y_1 - x_1) > 0.$$ 

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 $1-1$,

$$\left(\phi(y_2) - \phi(x_2)\right)(y_2 - x_2) < 0.$$ 

Let $x_t = tx_1 + (1-t)x_2$ and $y_t = ty_1 + (1-t)y_2$. Then $x_t < y_t$ for all $t \in [0, 1]$ because

$$tx_1 \leq ty_1$$

with strict inequality holding for at least one of these inequalities since not both $t$ and $(1-t)$ can equal zero. Now define

$$h(t) = (\phi(y_t) - \phi(x_t))(y_t - x_t).$$

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 $a$,

$$\phi(a) = \lim_{x \to a^+} \phi(z) \leq \phi(y) < \phi(x).$$

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 14.2.10 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 $y = f^{-1}(f(y)) < x = f^{-1}(f(x))$ 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 $\delta = \min(|f(x) - f(x + \eta), f(x - \eta) - f(x))$. 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 14.2.11** 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, \((f^{-1})'(f(x_1)) = \frac{1}{f'(x_1)}\).

**Proof:** By Lemma 14.2.10, \( f \) is either strictly increasing or strictly decreasing and \( f^{-1} \) is continuous on \([a, b]\). Therefore there exists \( \eta > 0 \) such that if \( 0 < |f'(x_1)| < \eta \), 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 \( 0 < |f'(x_1)| < \eta \),
\[ \left| \frac{f^{-1}(f(x)) - f^{-1}(f(x_1))}{f(x) - f(x_1)} - \frac{1}{f'(x_1)} \right| = \left| \frac{x - x_1}{f(x) - f(x_1)} - \frac{1}{f'(x_1)} \right| < \varepsilon \]
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 14.2.12** 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 in remembering formulas such as the one given in the theorem.
14.2. INDEPENDENCE OF PARAMETRIZATION

14.2.2 Independence Of Parametrization

Here is the precise definition of what is meant by a smooth curve.

**Definition 14.2.13** $C$ is a smooth curve in $\mathbb{R}^n$ if there exists an interval, $[a,b] \subseteq \mathbb{R}$ and functions $x_i : [a,b] \to \mathbb{R}$ such that the following conditions hold

1. $x_i$ is continuous on $[a,b]$.
2. $x_i'$ 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 $p (t) \equiv (x_1(t), \ldots, x_n(t))$, $t \to p (t)$ is one to one on $(a,b)$.
4. $|p'(t)| \equiv \left( \sum_{i=1}^{n} |x_i'(t)|^2 \right)^{1/2} \neq 0$ for all $t \in [a,b]$.
5. $C = \cup \{(x_1(t), \ldots, x_n(t)) : t \in [a,b]\}$.

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 parametrization 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 parametrization $(x_1, \ldots, x_n)$.

**Theorem 14.2.14** 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\(^\text{\textsuperscript{\textregistered}}\),

\[ \int_a^b f(s) \, ds = \int_a^c f(\phi(t)) \left| \phi'(t) \right| \, dt \]

**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 \(^\text{\textsuperscript{\textregistered}}\), $\phi$ 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 but easier. This proves the theorem.

**Lemma 14.2.15** Let $f : [a,b] \to C$, $g : [c,d] \to C$ be parameterizations of a smooth curve which satisfy conditions \(^\text{\textsuperscript{\textregistered}}\). 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 $[a,b]$.

\(^\text{\textsuperscript{\textregistered}}\)Recall that all continuous functions of this sort are Riemann integrable.
The length of a smooth curve is not dependent on parametrization.

**Proof**: It is obvious φ 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 14.2.16. If φ is not continuous on [a, b], then there exists a sequence, \( \{t_n\} \subseteq [a, b] \) such that \( t_n \to t \) but φ(t_n) fails to converge to φ(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 14.2.16 on Page 294), 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 14.2.16**: The length of a smooth curve is not dependent on parametrization.

**Proof**: Let \( C \) be the curve and suppose \( f : [a, b] \to C \) and \( g : [c, d] \to C \) both satisfy conditions 14.2.1. 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 \( s_0 \in \phi([a + \delta, b - \delta]) \subseteq (c, d) \). Then by assumption 14.2, \( g_i'(s_0) \neq 0 \) 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 \( s \in I \),

\[
h_i'(g_i(s)) = \frac{1}{g_i'(s)}.
\]

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 \( t \in J_1 \),

\[
\phi'(t) = h_i'(f_i(t)) f_i'(t) = h_i'(g_i(s)) f_i'(t) = \frac{f_i'(t)}{g_i'(\phi(t))} \tag{14.4}
\]

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 14.2.13,

\[
\int_{\phi(a + \delta)}^{\phi(b - \delta)} |g'(s)| \, ds = \int_{a + \delta}^{b - \delta} |g'(\phi(t))| \, |\phi'(t)| \, dt = \int_{a + \delta}^{b - \delta} |f'(t)| \, dt.
\]
14.3 Product Rule For Matrices

Another kind of multiplication is matrix multiplication. Here is the concept of the product rule extended to matrix multiplication.

**Definition 14.3.1** 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 \( A'(t) = (A'_{ij}(t)) \). 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 14.3.2** 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.

14.4 Moving Coordinate Systems

Let \( \mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t) \) be a right handed 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 \( \mathbf{i}(t) \), 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. Now let \( \mathbf{u} = (u_1, u_2, u_3) \in \mathbb{R}^3 \) and let \( t_0 \) be some reference time. For example you could let \( t_0 = 0 \). Then define the components of \( \mathbf{u} \) with respect to these vectors, \( \mathbf{i}, \mathbf{j}, \mathbf{k} \) at time \( t_0 \) as

\[
\mathbf{u} \equiv u_1 \mathbf{i}(t_0) + u_2 \mathbf{j}(t_0) + u_3 \mathbf{k}(t_0).
\]

Let \( \mathbf{u}(t) \) be defined as the vector which has the same components with respect to \( \mathbf{i}, \mathbf{j}, \mathbf{k} \) but at time \( t \). Thus

\[
\mathbf{u}(t) \equiv u_1 \mathbf{i}(t) + u_2 \mathbf{j}(t) + u_3 \mathbf{k}(t).
\]

and the vector has changed although the components have not.

\(^3\)Recall that right handed implies \( \mathbf{i} \times \mathbf{j} = \mathbf{k} \).
Suppose, vector, \(Q\)

**Lemma 14.4.1**

Recall this means you take a vector, \(u\) with respect to \(i\), \(u\) and using the repeated index summation convention,

\[
\text{showing that } Q(t) \text{ is a linear transformation. Also, } Q(t) \text{ preserves all distances because, since the vectors, } i(t), j(t), k(t) \text{ form an orthonormal set,}
\]

\[
|Q(t)u| = \left( \sum_{i=1}^{3} (u_i)^2 \right)^{1/2} = |u|.
\]

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 \(i(t_0), j(t_0), k(t_0)\). I will refer to this matrix as \(Q(t)\) to save notation.

**Lemma 14.4.1** 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,

\[
(Q(t)u)(Q(t)w) = \frac{1}{4} \left( (Q(t)(u + w))^2 - |Q(t)(u - w)|^2 \right)
\]

\[
= \frac{1}{4} \left( |u + w|^2 - |u - w|^2 \right)
\]

\[
= (u \cdot w).
\]
This implies
\[
(Q(t)^T Q(t) \mathbf{u} \cdot \mathbf{w}) = (\mathbf{u} \cdot \mathbf{w})
\]
for all \(\mathbf{u}, \mathbf{w}\). Therefore, \(Q(t)^T Q(t) \mathbf{u} = \mathbf{u}\) and so \(Q(t)^T Q(t) = Q(t)Q(t)^T = I\). This proves the first part of the lemma.

It follows from the product rule, Lemma 14.3.2 that
\[
Q'(t)Q(t)^T + Q(t)Q'(t)^T = 0
\]
and so
\[
Q'(t)Q(t)^T = - \left( Q'(t)Q(t)^T \right)^T.
\] (14.5)

From the definition, \(Q(t)\mathbf{u} = \mathbf{u}(t)\),
\[
\mathbf{u}'(t) = Q'(t)\mathbf{u} = \frac{\partial}{\partial t} \left( Q(t)^T \mathbf{u}(t) \right).
\]

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 Theorem 14.4.1 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,
\[
\begin{pmatrix}
u_1 \\
u_2 \\
u_3
\end{pmatrix}'(t) = \begin{pmatrix}
0 & -\omega_3(t) & \omega_2(t) \\
\omega_3(t) & 0 & -\omega_1(t) \\
-\omega_2(t) & \omega_1(t) & 0
\end{pmatrix} \begin{pmatrix}
u_1 \\
u_2 \\
u_3
\end{pmatrix}(t)
= \begin{pmatrix}
w_2(t)u_3(t) - w_3(t)u_2(t) \\
w_3(t)u_1(t) - w_1(t)u_3(t) \\
w_1(t)u_2(t) - w_2(t)u_1(t)
\end{pmatrix}
\]
where the \(u_i\) are the components of the vector \(\mathbf{u}(t)\) in terms of the fixed vectors \(i(t_0), j(t_0), k(t_0)\).

Therefore,
\[
\mathbf{u}'(t) = \boldsymbol{\Omega}(t) \times \mathbf{u}(t) = Q'(t)Q(t)^T \mathbf{u}(t)
\] (14.6)
where
\[
\boldsymbol{\Omega}(t) = \omega_1(t)i(t_0) + \omega_2(t)j(t_0) + \omega_3(t)k(t_0).
\]

because
\[
\boldsymbol{\Omega}(t) \times \mathbf{u}(t) \equiv \begin{vmatrix}
i(t_0) & j(t_0) & k(t_0) \\
w_1 & w_2 & w_3 \\
u_1 & u_2 & u_3
\end{vmatrix} \equiv
\begin{vmatrix}
i(t_0)(w_2u_3 - w_3u_2) + j(t_0)(w_3u_1 - w_1u_3) + k(t_0)(w_1u_2 - w_2u_1)
\end{vmatrix}
\]
\[
\text{This proves the lemma and yields the existence part of the following theorem.}
\]

Theorem 14.4.2 Let \(i(t), j(t), k(t)\) be as described. Then there exists a unique vector \(\boldsymbol{\Omega}(t)\) such that if \(\mathbf{u}(t)\) is a vector whose components are constant with respect to \(i(t), j(t), k(t)\), then
\[
\mathbf{u}'(t) = \boldsymbol{\Omega}(t) \times \mathbf{u}(t).
\]
CHAPTER 14. PHYSICS OF CURVILINEAR MOTION

Proof: It only remains to prove uniqueness. Suppose \( \Omega_1 \) also works. Then \( u(t) = Q(t)u \) and so \( u'(t) = Q'(t)u \) and

\[
Q'(t)u = \Omega \times Q(t)u = \Omega_1 \times Q(t)u
\]

for all \( u \). Therefore,

\[
(\Omega - \Omega_1) \times Q(t)u = 0
\]

for all \( u \) and since \( Q(t) \) is one to one and onto, this implies \( (\Omega - \Omega_1) \times w = 0 \) for all \( w \) and thus \( \Omega - \Omega_1 = 0 \). This proves the theorem.

Definition 14.4.3 A rigid body in \( \mathbb{R}^3 \) has 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 \( u(t) \) 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 \( u(t) \) is fixed with the body if \( u(t) = u_1 \mathbf{i}(t) + u_2 \mathbf{j}(t) + u_3 \mathbf{k}(t) \).

The following comes from the above discussion.

Theorem 14.4.4 Let \( B(t) \) be the set of points in three dimensions occupied by a rigid body. Then there exists a vector \( \Omega(t) \) such that whenever \( u(t) \) is fixed with the rigid body,

\[
u'(t) = \Omega(t) \times u(t).
\]
Part IV

Functions Of Many Variables
Chapter 15

Functions Of Many Variables

Quiz

1. Let \( r(t) = (\cos(t), \sin(t), 2t) \). Find \( a_T \), and \( a_N \). Also find \( \kappa \) and write the acceleration as the sum of two terms, one in the direction of the unit tangent vector and the other in the direction of the principle normal. Find \( \vec{\kappa} \), the curvature vector which is a completely useless concept.

2. Here is a matrix which happens to have \(-1\) as an eigenvalue. Find the eigenspace corresponding to this eigenvalue.

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

Is the matrix defective or nondefective?

3. In Problem 2 find the determinant of the matrix.

4. Raise the matrix of Problem 2 to the 15th power exactly.

15.1 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.

\[ \]

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 15.1.1** Determine from a contour map where the function, \( f(x, y) = \sin(x^2 + y^2) \) is steepest.
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.

15.2 The Domain Of A Function

As usual the domain of a function is either specified or if it is unspecified, it is the set of all points for which the function makes sense. If $f$ is the name of the function its domain is denoted as $D(f)$.

**Example 15.2.1** Find the domain of the function, $f(x, y) = \sqrt{1 - (x^2 + y^2)}$.

You need to have $1 \geq x^2 + y^2$ and so the domain of this function is $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$. This is just the inside of the unit circle centered at $(0, 0)$. It also includes the edge of this unit circle.

Sometimes the domain is given to you in a very artificial way.

**Example 15.2.2** Let $D(f) = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\} \cup (3, 7)$. Let $f(x, y) = x + 2y$ for $(x, y) \in \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$ and let $f(3, 7) = 33$.

In this case, the domain of the function is as given above and the function is given the definition just described.

15.3 Quadric Surfaces

**Quiz**

1. Find the eigenvectors of the matrix,

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

given the eigenvalues are 6 and 3. Also tell whether the matrix is defective or non defective.
2. Here is a matrix.

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

Find \( A^{50} \). The eigenvalues of this matrix are 1 and 1/2 and eigenvectors for these eigenvalues are \( \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \) for \( \lambda = 1 \) and \( \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \) and \( \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \) for \( \lambda = 1/2 \).

3. Here is a Markov matrix. This is also called a migration matrix or a stochastic matrix.

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

Find

\[ \lim_{n \to \infty} A^n \begin{pmatrix} 16 \\ 30 \\ 5 \end{pmatrix}. \]

Recall the equation of an arbitrary plane is an equation of the form \( ax + by + cz = d \).

More generally, a set of points of the following form

\[ \{(x, y, z) : f(x, y, z) = c\} \]

is called a level surface. 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.

\[ \frac{z^2}{a^2} - \frac{x^2}{b^2} - \frac{y^2}{c^2} = 1 \]

hyperboloid of two sheets

\[ \frac{x^2}{b^2} + \frac{y^2}{c^2} - \frac{z^2}{a^2} = 1 \]

hyperboloid of one sheet
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, $z = \frac{x^2}{a^2} - \frac{y^2}{b^2}$. If $z > 0$ this is the equation of a hyperbola which opens to the right and left while if $z < 0$ it is a hyperbola.

\footnote{It is traditional to refer to this as a hyperbolic paraboloid. Not a parabolic hyperboloid.}
CHAPTER 15. FUNCTIONS OF MANY VARIABLES

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 15.3.1** 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 15.3.2** 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 15.3.3** 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 15.3.4** Find the points of intersection of the line \( (x, y, z) = (1 + t, 1 + 2t, 1 + t) \) with the surface, \( z = x^2 + y^2 \).

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

\[
\left(\frac{1}{2} + \frac{1}{10} \sqrt{5}, \frac{1}{5} \sqrt{5}, \frac{1}{2} + \frac{1}{10} \sqrt{5}\right), \quad \left(\frac{1}{2} - \frac{1}{10} \sqrt{5}, -\frac{1}{5} \sqrt{5}, \frac{1}{2} - \frac{1}{10} \sqrt{5}\right).
\]

A cylinder generated by a curve, \( C \) is the surface generated by moving the curve \( C \) through space along a straight line. If you are given a level surface of the form \( f(x, y) = c \) this will yield a cylinder parallel to the \( z \) axis. Here is why: If \( z = 0 \), then \( f(x, y) = c \) is a curve in the plane, \( z = 0 \). If \( z = 1 \), then you get exactly the same curve but just shifted up to a height of 1. Similarly, \( f(y, z) = c \) gives a cylinder parallel to the \( x \) axis and \( f(x, z) = c \) gives one which is parallel to the \( y \) axis.

**Example 15.3.5** Consider the cylinder \( x^2 + y^2 = 1 \). Sketch its graph.
You see that at every height above or below the \( z = 0 \) plane if you slice it at that level, you will just see the graph of \( x^2 + y^2 = 1 \) at that level. This is a circle of radius 1. Since the equation describing the surface does not depend on \( z \), this is why it looks the same at every level.

### 15.4 Open And Closed Sets

Now remember from calculus of functions of one variable some of the things you did. One of the most important was to consider the derivative of a function. Recall the definition of the derivative, \( f'(x) \).

\[
\lim_{y \to x} \frac{f(y) - f(x)}{y - x} = f'(x).
\]

In order to write this definition you need to have \( f \) defined for all values of \( y \) near \( x \), That is, you need to have \( f \) defined on an open interval containing \( x \) of the form \( (x - \delta, x + \delta) \) for some \( \delta > 0 \). Otherwise, you can’t consider \( f(y) \). This is why the concepts to be discussed here are so important.

We are going to consider functions 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 \( x \). It is essential you understand a few kinds of sets.

**Definition 15.4.1** Let \( x \in \mathbb{R}^n \). Then \( B(x, r) \), called the ball centered at \( x \) having radius \( r \) is defined to be the set of all points of \( \mathbb{R}^n \), \( y \) which have the property that these points are closer than \( r \) to \( x \). Thus \( y \in B(x, r) \) means \( |y - x| < r \). Written formally, this is

\[
B(x, r) \equiv \{ y \in \mathbb{R}^n : |y - x| < r \}.
\]

To say that \( B(x, r) \subset D(f) \) means that whenever \( y \) is closer to \( x \) than \( r \), it follows \( y \in D(f) \). Now recall this is the sort of thing which you must start with, even in one dimension, to consider the concept of the derivative of a function. Therefore, it is not surprising that such an idea would be important in \( \mathbb{R}^n \).
Definition 15.4.2 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 15.4.3 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 $\mathbb{R}^n \setminus S$. A bounded set is one which is contained in a large enough ball. In $\mathbb{R}^n$ a set which is both closed and bounded is compact.\footnote{Actually the term compact has independent meaning and there is a theorem called the Heine Borel theorem which states that in $\mathbb{R}^n$ closed and bounded sets are compact. See the section on theory for more on this. This is not just useless jargon and gratuitous terminology.}

To illustrate this definition, consider the following picture.

\begin{center}
\begin{tikzpicture}
  \node at (0,0) [circle,fill,inner sep=1.5pt] (x) {};
  \node at (0,0) [circle,dotted] (B) {$B(x,r)$};
  \node at (0,0) [circle,dotted] (U) {$U$};
\end{tikzpicture}
\end{center}

You see in this picture how the edges are dotted. This is because an open set, can’t 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. So why might it be important to consider closed sets? Remember from one variable calculus the theorem which says that a continuous function achieves its maximum and minimum on a closed interval. The closed interval contains its “skin”, the end points of the interval. Similar theorems will end up holding for functions of $n$ variables. Here is a picture of a closed set, $C$.\footnote{Actually the term compact has independent meaning and there is a theorem called the Heine Borel theorem which states that in $\mathbb{R}^n$ closed and bounded sets are compact. See the section on theory for more on this. This is not just useless jargon and gratuitous terminology.}
15.4. OPEN AND CLOSED SETS

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. 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 15.4.4** 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

---

*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.*
triangle inequality that
\[ |z - x| = |z - y + y - x| \leq |z - y| + |y - x| < r + |y - x| = r - |x - y| + |y - x| = r. \]

Note that if \( r = 0 \) then \( B(x, r) = \emptyset \), the empty set. This is because if \( y \in \mathbb{R}^n, |x - y| \geq 0 \) and so \( y \notin B(x, 0) \). Since \( \emptyset \) 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,
\[
|x - z| \geq |x - y| - |y - z| > |x - y| - \delta = |x - y| - (|x - y| - r) = r
\]

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.

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 15.4.5** The Cartesian product of two sets, \( A \times B \), means \( \{(a, b) : a \in A, b \in B \} \).

If you have \( n \) sets, \( A_1, A_2, \cdots, A_n \)

\[
\prod_{i=1}^{n} A_i = \{(x_1, x_2, \cdots, x_n) : \text{each } x_i \in A_i \}.
\]

Now suppose \( A \subseteq \mathbb{R}^m \) and \( B \subseteq \mathbb{R}^n \). Then if \((x, y) \in A \times B, x = (x_1, \cdots, x_m) \) and \( y = (y_1, \cdots, y_n) \), the following identification will be made.

\[
(x, y) = (x_1, \cdots, x_m, y_1, \cdots, y_n) \in \mathbb{R}^{n+m}.
\]

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 15.4.6** 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 15.4.7 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

$$B((x, y) , \delta) = \left\{ (s, t) \in \mathbb{R}^{m+n} : \sum_{k=1}^{m} |x_k - s_k|^2 + \sum_{j=1}^{n} |y_j - t_j|^2 < \delta^2 \right\}$$

Therefore, if $\delta = \min(r_1, r_2)$ 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 $B((x, y) , \delta) \not\subseteq \mathbb{R}^{m+n} \setminus (C \times H)$. 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, $B((x, y) , r) \subseteq \mathbb{R}^{m+n} \setminus (C \times H)$ 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 $C \subseteq \prod_{i=1}^{m} [a_i, b_i]$ 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.

15.5 Continuous Functions

What was done in beginning calculus for scalar functions is generalized here to include the case of a vector valued function of possibly many variables. What follows is the **correct definition** of continuity. The one you are used to seeing in terms of the value of the function corresponding to the value of its limit is not correct in general. This one you are used to seeing is only correct if the point of the domain of the function is a limit point of the domain, discussed briefly later (Don’t worry about it too much. Just use the correct definition and you will be fine.). It isn’t a big deal for functions of one variables because you usually are dealing with functions defined on intervals and it happens that all the points are limit points. In multiple dimensions, however, the earlier definition is woefully inadequate and will lead you to profound confusion, confusion which is so severe you will have to relearn everything you thought you understood. I know this from bitter personal experience.

**Definition 15.5.1** 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.
15.6 Sufficient Conditions For Continuity

The next theorem is a fundamental result which allows less worry about the $\varepsilon \delta$ definition of continuity.

**Theorem 15.6.1** 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.

The proof of this theorem is given later. 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 15.6.2** 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

$$|\alpha| \equiv \sum_{i=1}^{n} |\alpha_i|$$

The symbol, $x^\alpha$ means

$$x^\alpha \equiv x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}.$$ 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.
15.7 Properties Of Continuous Functions

Functions of many 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 15.7.1** 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 $x \in C$,

$$f(x_1) \leq f(x) \leq f(x_2).$$

The above theorems are proved in an optional section.
Chapter 16
Limits Of A Function

Quiz

1. The position vector of an object is \( r(t) = (e^t, \sin(t), t^2 - 1) \). Find the unit tangent vector when \( t = 0 \).

2. Show that for \( v(t) \) a vector valued function \( \frac{d}{dt} |v(t)| = \frac{v \cdot v}{|v|} \) (Note that in the case where \( v \) is velocity, this implies \( \mathbf{a} \cdot \mathbf{T} = \frac{d}{dt} |\mathbf{v}| \).

3. Suppose \( r(t) = (t^2, \cos(t), \sin(t)) \). Find the curvature when \( t = 0 \).

4. Suppose \( r(t) = (2t^{1/2}, \frac{2}{3}t^{3/2}, \sqrt{2}t) \) for \( t \in [1, 2] \). Find the length of this curve.

5. Find the matrix of the linear transformation which projects all vectors onto the line \( y = x \).

16.1 The Limit Of A Function Of Many Variables

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. It is a harder concept than the concept of continuity.

Definition 16.1.1 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 16.1.2 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

\[
\lim_{y \to x} f(y) = L
\]

if and only if the following condition holds. For all \( \varepsilon > 0 \) there exists \( \delta > 0 \) such that if

\[
0 < |y - x| < \delta, \text{ and } y \in D(f)
\]

then,

\[
|L - f(y)| < \varepsilon.
\]

Theorem 16.1.3 If \( \lim_{y \to x} f(y) = L \) and \( \lim_{y \to x} f(y) = L_1 \), then \( L = L_1 \).
Proof: Let $\varepsilon > 0$ be given. There exists $\delta > 0$ such that if $0 < |y - x| < \delta$ and $y \in D(f)$, then
\[ |f(y) - L| < \varepsilon, \ |f(y) - L_1| < \varepsilon. \]
Pick such a $y$. There exists one because $x$ is a limit point of $D(f)$. Then
\[ |L - L_1| \leq |L - f(y)| + |f(y) - L_1| < \varepsilon + \varepsilon = 2\varepsilon. \]
Since $\varepsilon > 0$ was arbitrary, this shows $L = L_1$.

As in the case of functions of one variable, one can define what it means for $\lim_{y\to x} f(x) = \pm \infty$.

Definition 16.1.4 If $f(x) \in \mathbb{R}$, $\lim_{y\to x} f(x) = \infty$ if for every number $l$, there exists $\delta > 0$ such that whenever $|y - x| < \delta$ and $y \in D(f)$, then $f(y) > l$.

The following theorem is just like the one variable version of calculus.

Theorem 16.1.5 Suppose $\lim_{y\to x} f(y) = L$ and $\lim_{y\to x} g(y) = K$ where $K, L \in \mathbb{R}^q$. Then if $a, b \in \mathbb{R}$,
\[ \lim_{y\to x} (af(y) + bg(y)) = aL + bK, \tag{16.1} \]
\[ \lim_{y\to x} f(y) \cdot g(y) = L \cdot K \tag{16.2} \]
and if $g$ is scalar valued with $\lim_{y\to x} g(y) = K \neq 0$,
\[ \lim_{y\to x} f(y) g(y) = LK. \tag{16.3} \]

Also, if $h$ is a continuous function defined near $L$, then
\[ \lim_{y\to x} h \circ f(y) = h(L). \tag{16.4} \]

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 (16.1) is left for you. It is like a corresponding theorem for continuous functions. Now (16.2) is to be verified. Let $\varepsilon > 0$ be given. Then by the triangle inequality,
\[ |f \cdot g(y) - L \cdot K| \leq |f(y) - f(y) \cdot K| + |f(y) \cdot K - L \cdot K| \leq |f(y)| |g(y) - K| + |K| |f(y) - L|. \]

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 $0 < |y - x| < \delta_1$,
\[ |f \cdot g(y) - L \cdot K| \leq (1 + |K| + |L|) |g(y) - K| + |f(y) - L|. \tag{16.5} \]
Now let $0 < \delta_2$ be such that if $y \in D(f)$ and $0 < |x - y| < \delta_2$,
\[ |f(y) - L| < \frac{\varepsilon}{2(1 + |K| + |L|)}, \ |g(y) - K| < \frac{\varepsilon}{2(1 + |K| + |L|)}. \]

Then letting $0 < \delta \leq \min (\delta_1, \delta_2)$, it follows from (16.7) that
\[ |f \cdot g (y) - L \cdot K| < \varepsilon \]
and this proves (16.6).

The proof of (16.3) is left to you.

Consider (16.5). Since $h$ is continuous near $L$, it follows that for $\varepsilon > 0$ given, there exists $\eta > 0$ such that if $|y - L| < \eta$, then
\[ |h (y) - h (L)| < \varepsilon \]
Now since $\lim_{y \to x} f (y) = L$, there exists $\delta > 0$ such that if $0 < |y - x| < \delta$, then
\[ |f (y) - L| < \eta. \]
Therefore, if $0 < |y - x| < \delta$,
\[ |h (f (y)) - h (L)| < \varepsilon. \]

It only remains to verify the last assertion. Assume $|f (y) - b| \leq r$. It is required to show that $|L - b| \leq r$. If this is not true, then $|L - b| > r$. Consider $B (L, |L - b| - r)$. Since $L$ is the limit of $f$, it follows $f (y) \in B (L, |L - b| - r)$ whenever $y \in D (f)$ is close enough to $x$. Thus, by the triangle inequality,
\[ |f (y) - L| < |L - b| - r \]
and so
\[ r < |L - b| - |f (y) - L| \leq ||b - L| - |f (y) - L|| \]
\[ \leq |b - f (y)|, \]
a contradiction to the assumption that $|b - f (y)| \leq r$.

The next theorem gives the correct relation between continuity and the limit.

**Theorem 16.1.6** For $f : D (f) \to \mathbb{R}^q$ and $x \in D (f)$ a limit point of $D (f)$, $f$ is continuous at $x$ if and only if
\[ \lim_{y \to x} f (y) = f (x). \]

**Proof:** First suppose $f$ is continuous at $x$ a limit point of $D (f)$. Then for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $|y - x| < \delta$ and $y \in D (f)$, then $|f (x) - f (y)| < \varepsilon$. In particular, this holds if $0 < |x - y| < \delta$ and this is just the definition of the limit. Hence $f (x) = \lim_{y \to x} f (y)$.

Next suppose $x$ is a limit point of $D (f)$ and $\lim_{y \to x} f (y) = f (x)$. This means that if $\varepsilon > 0$ there exists $\delta > 0$ such that for $0 < |x - y| < \delta$ and $y \in D (f)$, it follows $|f (y) - f (x)| < \varepsilon$. However, if $y = x$, then $|f (y) - f (x)| = |f (x) - f (x)| = 0$ and so whenever $y \in D (f)$ and $|x - y| < \delta$, it follows $|f (x) - f (y)| < \varepsilon$, showing $f$ is continuous at $x$.

The following theorem is important.

**Theorem 16.1.7** Suppose $f : D (f) \to \mathbb{R}^q$. Then for $x$ a limit point of $D (f)$,
\[ \lim_{y \to x} f (y) = L \]  \hspace{1cm} (16.6)
if and only if
\[ \lim_{y \to x} f_k (y) = L_k \]  \hspace{1cm} (16.7)
where $f (y) = (f_1 (y), \cdots , f_p (y))$ and $L = (L_1 , \cdots , L_p)$.

In the case where $q = 3$ and $\lim_{y \to x} f (y) = L$ and $\lim_{y \to x} g (y) = K$, then
\[ \lim_{y \to x} f (y) \times g (y) = L \times K. \]  \hspace{1cm} (16.8)
CHAPTER 16. LIMITS OF A FUNCTION

**Proof:** Suppose \(16.6\). Then letting \(\varepsilon > 0\) be given there exists \(\delta > 0\) such that if \(0 < |y - x| < \delta\), it follows

\[
|f_k(y) - L_k| \leq |f(y) - L| < \varepsilon
\]

which verifies \(16.7\).

Now suppose \(16.7\) holds. Then letting \(\varepsilon > 0\) be given, there exists \(\delta_k\) such that if \(0 < |y - x| < \delta_k\), then

\[
|f_k(y) - L_k| < \frac{\varepsilon}{\sqrt{p}}.
\]

Let \(0 < \delta < \min(\delta_1, \ldots, \delta_p)\). Then if \(0 < |y - x| < \delta\), it follows

\[
|f(y) - L| = \left(\sum_{k=1}^{p} |f_k(y) - L_k|^2\right)^{1/2} < \left(\sum_{k=1}^{p} \frac{\varepsilon^2}{\sqrt{p}}\right)^{1/2} = \varepsilon.
\]

It remains to verify \(16.8\). But from the first part of this theorem and the description of the cross product presented earlier in terms of the permutation symbol,

\[
\lim_{y \to x} (f(y) \times g(y))_i = \lim_{y \to x} \varepsilon_{ijk} f_j(y) g_k(y) = \varepsilon_{ijk} L_j K_k = (L \times K)_i.
\]

If you did not read about the permutation symbol, you can simply write out the cross product and observe that the desired limit holds for each component. Therefore, from the first part of this theorem, this establishes \(16.8\). This completes the proof.

**Example 16.1.8** Find \(\lim_{(x,y) \to (3,1)} \left(\frac{x^2-9}{x-3}, y\right)\).

It is clear that \(\lim_{(x,y) \to (3,1)} \frac{x^2-9}{x-3} = 6\) and \(\lim_{(x,y) \to (3,1)} y = 1\). Therefore, this limit equals \((6, 1)\).

**Example 16.1.9** 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.
16.2 The Directional Derivative And Partial Derivatives

16.2.1 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 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

\[
D_{\mathbf{v}} f(x) \equiv \lim_{t \to 0} \frac{f(x_0 + tv_1, y_0 + tv_2) - f(x_0, y_0)}{t}.
\]

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 16.2.1** 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 16.2.2** Find the directional derivative of the function, \( f(x, y) = x^2y \) 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 \( f(x + tv) = \left(1 + \frac{t\sqrt{2}}{2}\right)^2 \left(2 + \frac{t\sqrt{2}}{2}\right) \) and \( f(x) = 2 \). Therefore,

\[
\frac{f(x + tv) - f(x)}{t} = \left(1 + \frac{t\sqrt{2}}{2}\right)^2 \left(2 + \frac{t\sqrt{2}}{2}\right) - 2,
\]

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 \( t \to 0 \),

\[
D_v f (1, 2) = \left(\frac{5}{2}\sqrt{2}\right).
\]

There is something you must keep in mind about this. The direction vector must always be a unit vector.\(^4\)

### 16.2.2 Partial Derivatives

**Quiz**

1. Let \( r(t) = (t^2, \cosh(t) + t, \sin(t)) \). Find \( \kappa \) when \( t = 0 \). Remember \( \kappa \) is the curvature. Also find the normal and tangential components of acceleration and the osculating plane at the point where \( t = 0 \).

2. Suppose \( |r(t)| = 33 \) for all \( t \). Show that \( r'(t) \cdot r(t) = 0 \). Does it follow that \( r'(t) = 0? \)

3. Suppose \( r(t) = (2t^{1/2}, \frac{2}{7}t^{3/2}, \sqrt{2}t) \) for \( t \in [1, 2] \). Find the length of this curve.

There are some special unit vectors which come to mind immediately. These are the vectors, \( e_i \), where

\[
e_i = (0, \cdots, 0, 1, 0, \cdots, 0)^T
\]

and the 1 is in the \( i^{th} \) position.

Thus in case of a function of two variables, the directional derivative in the direction \( i = e_1 \) is the slope of the indicated straight line in the following picture.

\(^4\)Actually, there is a more general formulation of the notion of directional derivative known as the Gateaux derivative in which the length of \( v \) is not one but it is not considered here.
16.2. THE DIRECTIONAL DERIVATIVE AND PARTIAL DERIVATIVES

As in the case of a general directional derivative, you fix $y$ and take the derivative of the function, $x \to f(x, y)$. More generally, even in situations which cannot be drawn, the definition of a partial derivative is as follows.

**Definition 16.2.3** Let $U$ be an open subset of $\mathbb{R}^n$ and let $f : U \to \mathbb{R}$. Then letting $\mathbf{x} = (x_1, \ldots, x_n)^T$ be a typical element of $\mathbb{R}^n$,

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) \equiv D_i f(\mathbf{x}).$$

This is called the **partial derivative** of $f$. Thus,

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + te_i) - f(\mathbf{x})}{t} = \lim_{t \to 0} \frac{f(x_1, \ldots, x_i + t, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{t},$$

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 16.2.4** Find $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y},$ and $\frac{\partial f}{\partial z}$ if $f(x, y) = y \sin x + x^2 y + z$.

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.

**Definition 16.2.5** 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
\[
\nabla f(x) = (f_1(x), f_2(x), \ldots, f_n(x))^T.
\]

**Proposition 16.2.6** In the situation of Definition 16.2.5 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 16.2.7** Find the directional derivative of the function, \( f(x,y) = \sin (2x^2 + y^3) \) at \((1,1)\) in the direction \( (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) \). \( T \).

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 16.2.8** In the situation of Definition 16.2.5, 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 \(-|\nabla f(x)|\).
16.2. THE DIRECTIONAL DERIVATIVE AND PARTIAL DERIVATIVES

Proof: Let \( \mathbf{v} \) be any unit vector. Then from Proposition 16.2.8,

\[
D_\mathbf{v} f (\mathbf{x}) = \nabla f (\mathbf{x}) \cdot \mathbf{v} = |\nabla f (\mathbf{x})| |\mathbf{v}| \cos \theta = |\nabla f (\mathbf{x})| \cos \theta
\]

where \( \theta \) is the included angle between these two vectors, \( \nabla f (\mathbf{x}) \) and \( \mathbf{v} \). Therefore, \( D_\mathbf{v} f (\mathbf{x}) \) 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 \( \mathbf{v} = \nabla f (\mathbf{x}) / |\nabla f (\mathbf{x})| \) and \( D_\mathbf{v} f (\mathbf{x}) = |\nabla f (\mathbf{x})| \). The second case occurs when \( \theta \) is \( \pi \) and in this case the two vectors point in opposite directions and the directional derivative equals \(-|\nabla f (\mathbf{x})|\).

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 16.2.9 Let \( f : U \rightarrow \mathbb{R}^p \) where \( U \) is an open set in \( \mathbb{R}^n \) and let \( \mathbf{v} \) be a unit vector. For \( \mathbf{x} \in U \), define the directional derivative of \( f \) in the direction, \( \mathbf{v} \), at the point \( \mathbf{x} \) as

\[
D_\mathbf{v} f (\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t}.
\]

Example 16.2.10 Let \( f(x, y) = (xy^2, yx)^T \). 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

\[
\lim_{t \to 0} \left( \left( x + t \left( \frac{1}{\sqrt{5}} \right) \right) \left( y + t \left( \frac{2}{\sqrt{5}} \right) \right)^2 - xy^2, \ (x + t \left( \frac{1}{\sqrt{5}} \right)) \left( y + t \left( \frac{2}{\sqrt{5}} \right) \right) - xy \right)
\]

\[
= \lim_{t \to 0} \left( \frac{4}{5} xy\sqrt{5} + \frac{4}{5} xt + \frac{1}{5} \sqrt{5} y^2 + \frac{4}{5} ty + \frac{4}{25} t^2 \sqrt{5}, \frac{2}{5} x\sqrt{5} + \frac{1}{5} y\sqrt{5} + \frac{2}{5} t \right)
\]

\[
= \left( \frac{4}{5} xy \sqrt{5} + \frac{1}{5} \sqrt{5} y^2, \frac{2}{5} x \sqrt{5} + \frac{1}{5} y \sqrt{5} \right).
\]

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 16.2.11 Find the partial derivative with respect to \( x \) of the function \( f(x, y, z, w) = (xy^2, z \sin(xy), z^3)^T \).

From the above definition, \( f_x(x, y, z) = D_1 f(x, y, z) = (y^2, z \cos(xy), z^3)^T \).

Example 16.2.12 Let \( f, g \) be two functions defined on an open subset of \( \mathbb{R}^3 \) which have partial derivatives. Find a formula for \( \nabla (fg) \).

This equals

\[
(fg)_x, (fg)_y, (fg)_z = (fxg + fgx, fyg + fgy, fzx + fgz)
\]

\[
= g(f_x, f_y, f_z) + f(g_x, g_y, g_z) = g \nabla f + f \nabla g
\]

Example 16.2.13 Let \( f, g \) be functions and \( a, b \) be scalars, you should verify that \( \nabla (af + bg) = a \nabla f + b \nabla g \).
Example 16.2.14 Let \( h ( x, y ) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \). Find \( \frac{\partial h}{\partial x} \) and \( \frac{\partial h}{\partial y} \).

If \( x > 0 \) or \( x < 0 \), both partial derivatives exist and equal 0. What of points like \((0, y)\)? \( \frac{\partial h}{\partial x} (0, y) \) does not exist but

\[
\frac{\partial h}{\partial y} (0, y) = \lim_{t \to 0} \frac{h(0, y + t) - h(0, 0)}{t} = \lim_{t \to 0} \frac{1 - 1}{t} = 0.
\]

Do not be afraid to use the definition of the partial derivatives. Sometimes it is the only way to find the partial derivative.

Example 16.2.15 Let \( u ( x, y ) = \ln (x^2 + y^2) \). Find \( u_{xx} + u_{yy} \).

First find \( u_x \). This equals \( \frac{2x}{x^2 + y^2} \). Next find \( u_{xx} \). This involves taking the partial derivative of \( u_x \). Thus it equals

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

Similarly \( u_{yy} = 2 \frac{x^2 - y^2}{(x^2 + y^2)^2} \) and so \( u_{xx} + u_{yy} = 0 \). Of course this assumes \( (x, y) \neq (0, 0) \).

16.2.3 Mixed Partial Derivatives

Under certain conditions the mixed partial derivatives will always be equal. The simple condition is that if they exist and are continuous, then they are equal. This astonishing fact is due to Euler in 1734. For reasons I cannot understand, calculus books hardly ever include a proof of this important result. It is not all that hard. Here it is.

Theorem 16.2.16 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

\[
\Delta(s, t) = \frac{1}{st} \left( f(x + t, y + s) - f(x + t, y) - f(x, y + s) - f(x, y) \right). \tag{16.9}
\]

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, \( h(t) = f(x + t, y + s) - f(x + t, y) \). Therefore, by the mean value theorem from calculus and the (one variable) chain rule,

\[
\Delta(s, t) = \frac{1}{st} (h(t) - h(0)) = \frac{1}{st} h'(\alpha t) t = \frac{1}{s} (f_x(x + \alpha t, y + s) - f_x(x + \alpha t, y))
\]
16.2. THE DIRECTIONAL DERIVATIVE AND PARTIAL DERIVATIVES

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 (16.9), \( \Delta (s, t) \) 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 16.2.17** 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 16.2.18**

\[
f(x, y) = \begin{cases} 
\frac{xy(x^2-y^2)}{x^2+y^2} & \text{if } (x, y) \neq (0, 0) \\
0 & \text{if } (x, y) = (0, 0)
\end{cases}
\]

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.

Here is a picture of the graph of this function. It looks innocuous but isn’t.
16.3 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. They are here to provide depth. If you want something more than a superficial knowledge, you should read this section. However, if you don’t want to deal with challenging topics, don’t read this stuff. Don’t even look at it.

**Theorem 16.3.1** 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 \(|f(x) - f(y)| < \frac{\varepsilon}{2(|a| + |b| + 1)}\) and there exists \(\delta_2 > 0\) such that whenever \(|x - y| < \delta_2\), it follows that \(|g(x) - g(y)| < \frac{\varepsilon}{2(|a| + |b| + 1)}\). Then let \(0 < \delta \leq \min(\delta_1, \delta_2)\). If \(|x - y| < \delta\), then everything happens at once. Therefore, using the triangle inequality

\[
|af(x) + bg(x) - (ag(y) + bg(y))| \\
\leq |a||f(x) - f(y)| + |b||g(x) - g(y)| \\
< |a| \left( \frac{\varepsilon}{2(|a| + |b| + 1)} \right) + |b| \left( \frac{\varepsilon}{2(|a| + |b| + 1)} \right) < \varepsilon.
\]

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

\[
|fg(x) - fg(y)| \leq \frac{\varepsilon}{2(1 + |g(x)| + |f(y)|)} \left( \frac{\varepsilon}{2(1 + |g(x)| + |f(y)|)} + \frac{\varepsilon}{2(1 + |g(x)| + |f(y)|)} \right) = \varepsilon.
\]

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 \(|x - y| < \min(\delta_0, \delta_1)\),

\[
\frac{|f(x) - f(y)|}{g(x) - g(y)} = \frac{\left| \frac{f(x)g(y) - f(y)g(x)}{g(x)g(y)} \right|}{\left| \frac{g(x) + f(x) - f(y)g(x)}{g(x)} \right|}
\]

\[
\leq \frac{2}{|g(x)|^2} \left| \frac{f(x)g(y) - f(y)g(x) + f(y)g(y) - f(y)g(x)}{2} \right|
\]

\[
\leq \frac{2}{|g(x)|^2} \left| \frac{3}{2} |g(x)||f(x) - f(y)| + |f(y)||g(y) - g(x)| \right|
\]

\[
\leq \frac{2}{|g(x)|^2} \left( 1 + 2|f(x)| + 2|g(x)| \right)||f(x) - f(y)| + |g(y) - g(x)||
\]

\[
= M \left| |f(x) - f(y)| + |g(y) - g(x)| \right|
\]

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 \(0 < \delta \leq \min (\delta_0, \delta_1, \delta_2, \delta_3)\), and \(|x - y| < \delta\), everything holds and

\[
\frac{|f(x) - f(y)|}{g(y)} \leq M \left[|f(x) - f(y)| + |g(y) - g(x)|\right]
\leq M \left[\frac{\varepsilon}{2} M^{-1} + \frac{\varepsilon}{2} M^{-1}\right] = \varepsilon.
\]

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 it is 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

\[
|f_k(x) - f_k(y)| \leq |f(x) - f(y)|
\]

\[
\equiv \left(\sum_{i=1}^{q} |f_i(x) - f_i(y)|^2\right)^{1/2}
\leq \sum_{i=1}^{q} |f_i(x) - f_i(y)|. \tag{16.10}
\]

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 (16.11) implies

\[
|f(x) - f(y)| \leq \sum_{i=1}^{q} |f_i(x) - f_i(y)|
\leq \sum_{i=1}^{q} \frac{\varepsilon}{q} = \varepsilon.
\]

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.
16.3. SOME FUNDAMENTALS*

16.3.1 The Nested Interval Lemma*

Here is a multidimensional version of the nested interval lemma.

Lemma 16.3.2 Let \( I_k = \prod_{i=1}^{p} [a_{i,k}, b_{i,k}] \equiv \{ x \in \mathbb{R}^p : x_i \in [a_{i,k}, b_{i,k}] \} \) and suppose that for all \( k = 1, 2, \ldots \),
\[
I_k \supseteq I_{k+1}.
\]
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 \( i = 1, \cdot \cdot \cdot , p \) , \([a_{i,k}, b_{i,k}] \supseteq [a_{i,k+1}, b_{i,k+1}]\).
This implies that for each \( i \),
\[
a_{i,k} \leq a_{i,k+1}, b_{i,k} \geq b_{i,k+1}.
\]
Consequently, if \( k \leq l \),
\[
a_{i,k} \leq b_{i,k} \leq b_{i,l}.
\]
Now define
\[
c_i \equiv \sup \{ a_{i,l} : l = 1, 2, \cdot \cdot \cdot \}
\]
By the first inequality in (16.14),
\[
c_i = \sup \{ a_{i,l} : l = k, k + 1, \cdot \cdot \cdot \}
\]
for each \( k = 1, 2 \cdot \cdot \cdot \). Therefore, picking any \( k \) shows that \( b_{i,k} \) is an upper bound for the set, \( \{ a_{i,l} : l = k, k + 1, \cdot \cdot \cdot \} \) 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 (16.15). Thus, for each \( i \) and each \( k \),
\[
a_{i,k} \leq c_i \leq b_{i,k}.
\]
Defining \( c \equiv (c_1, \cdot \cdot \cdot , c_p) \), \( c \in I_k \) 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 16.3.3 Let \( I_k = \prod_{i=1}^{p} [a_{i,k}, b_{i,k}] \equiv \{ x \in \mathbb{R}^p : x_i \in [a_{i,k}, b_{i,k}] \} \) and suppose that for all \( k = 1, 2, \cdot \cdot \cdot \),
\[
I_k \supseteq I_{k+1}.
\]
Then there exists a point, \( c \in \mathbb{R}^p \) which is an element of every \( I_k \).

Proof: For each \( i = 1, \cdot \cdot \cdot , p \) , \([a_{i,k}, b_{i,k}] \supseteq [a_{i,k+1}, b_{i,k+1}]\) and so by the nested interval theorem for one dimensional problems, there exists a point \( c_i \in [a_{i,k}, b_{i,k}] \) for all \( k \). Then letting \( c \equiv (c_1, \cdot \cdot \cdot , c_p) \) it follows \( c \in I_k \) for all \( k \). This proves the lemma.

16.3.2 The Extreme Value Theorem*

Definition 16.3.4 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
\[
\text{diam} (S) \equiv \sup \{ |x - y| : x, y \in S \}.
\]
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 \( \text{diam} (S) \) is just a careful description of what you would think of as the diameter. It measures how stretched out the set is.
Lemma 16.3.5 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 $C \subseteq \prod_{i=1}^p [a_i, b_i] = I_0$ 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 $[a_i, \frac{a_i + b_i}{2}]$ or $[c_i, b_i]$. 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 $\operatorname{diam} (I_0) = \left( \sum_{i=1}^p (b_i - a_i)^2 \right)^{1/2}$ because for $x, y \in I_0$, $|x_i - y_i| \leq |a_i - b_i|$ for each $i = 1, \cdots, p$,

$$|x - y| = \left( \sum_{i=1}^p |x_i - y_i|^2 \right)^{1/2} \leq 2^{-1} \left( \sum_{i=1}^p (b_i - a_i)^2 \right)^{1/2} = 2^{-1} \operatorname{diam} (I_0).$$

Denote by $\{J_1, \cdots, J_{2^p}\}$ these sets determined above. It follows the diameter of each set is no larger than $2^{-1} \operatorname{diam} (I_0)$. In particular, since $d \equiv (d_1, \cdots, d_p)$ and $c \equiv (c_1, \cdots, c_p)$ are two such points, for each $J_k$,

$$\operatorname{diam} (J_k) \equiv \left( \sum_{i=1}^p (d_i - c_i)^2 \right)^{1/2} \leq 2^{-1} \operatorname{diam} (I_0)$$

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 $x, y \in I_2$,

$$|x - y| \leq 2^{-1} \operatorname{diam} (I_1) \leq 2^{-2} \operatorname{diam} (I_2),$$

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 $\operatorname{diam} (I_k) \leq 2^{-k} \operatorname{diam} (I_0)$ 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 $\operatorname{diam} (I_0) 2^{-k} < r$. Then since $c \in I_k$, and any two points of $I_k$ are closer than $\operatorname{diam} (I_0) 2^{-k}$, $I_k$ 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 16.3.6** 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 16.3.5, \( M \) is a finite number. Is \( f(x_2) = M \) for some \( x_2 \)? if not, you could consider the function,

\[ g(x) = \frac{1}{M - f(x)} \]

and \( g \) would be a continuous and unbounded function defined on \( C \), contrary to Lemma 16.3.5. 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.

16.3.3 Sequences And Completeness*

Definition 16.3.7 A function whose domain is defined as a set of the form

\[ \{ k, k + 1, k + 2, \ldots \} \]

for \( k \) an integer is known as a sequence. Thus you can consider \( f(k), f(k + 1), f(k + 2), \ldots \), etc. Usually the domain of the sequence is either \( \mathbb{N} \), the natural numbers consisting of \( \{ 1, 2, 3, \ldots \} \) or the nonnegative integers, \( \{ 0, 1, 2, 3, \ldots \} \). Also, it is traditional to write \( f_1, f_2, \ldots \) instead of \( f(1), f(2), f(3), \ldots \) 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 \( f(k), f(k + 1), f(k + 2), \ldots \) but as \( \{ f_i \}_{i=k}^{\infty} \) 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 \( a_k \) 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 16.3.8 Let \( \{ a_k \}_{k=1}^{\infty} \) 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 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 \( a_1, \ldots, a_n \). This rule which specifies \( a_{n+1} \) from knowledge of \( a_k \) for \( k \leq n \) is known as a recurrence relation.

Example 16.3.9 Let \( a_1 = 1 \) and \( a_2 = 1 \). Assuming \( a_1, \ldots, a_{n+1} \) 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, \ldots. This particular sequence is called the Fibonacci sequence and is important in the study of reproducing rabbits.

Example 16.3.10 Let \( a_k = (k, \sin(k)) \). Thus this sequence has values in \( \mathbb{R}^2 \).

Definition 16.3.11 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 \( \{ a_n \} \).
For example, suppose \( a_n = (n^2 + 1) \). Thus \( a_1 = 2, a_3 = 10 \), etc. If
\[
n_1 = 1, n_2 = 3, n_3 = 5, \ldots, n_k = 2k - 1,
\]
then letting \( b_k = a_{n_k} \), it follows
\[
b_k = \left( (2k - 1)^2 + 1 \right) = 4k^2 - 4k + 2.
\]

**Definition 16.3.12** A sequence, \( \{a_k\} \) is said to converge to \( a \) if for every \( \epsilon > 0 \) there exists \( n_\epsilon \) such that if \( n > n_\epsilon \), then \( |a_n - a| < \epsilon \). 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 16.3.13** If a sequence, \( \{a_n\} \) converges to \( a \) and to \( b \) then \( a = b \).

**Proof:** There exists \( n_\epsilon \) such that if \( n > n_\epsilon \) then \( |a_n - a| < \frac{\epsilon}{2} \) and if \( n > n_\epsilon \), then \( |a_n - b| < \frac{\epsilon}{2} \). Then pick such an \( n \).
\[
|a - b| < |a - a_n| + |a_n - b| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\]
Since \( \epsilon \) is arbitrary, this proves the theorem.

The following is the definition of a Cauchy sequence in \( \mathbb{R}^p \).

**Definition 16.3.14** \( \{a_n\} \) is a Cauchy sequence if for all \( \epsilon > 0 \), there exists \( n_\epsilon \) such that whenever \( n, m \geq n_\epsilon \),
\[
|a_n - a_m| < \epsilon.
\]

A sequence is Cauchy means the terms are “bunching up to each other” as \( m, n \) get large.

**Theorem 16.3.15** 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 \( \epsilon = 1 \) 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 16.3.16** 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 \( k > K \), it follows that \( n_k > K \) because \( n_1, n_2, n_3, \ldots \) is strictly increasing as the subscript increases. Also, there exists \( K_1 \) such that if \( k, l \geq K_1 \), \( |a_n - a| < \frac{\varepsilon}{2} \). Then letting \( n > \max(K, K_1) \), pick \( k > \max(K, K_1) \).

Then

\[
|a - a_n| \leq |a - a_{n_k}| + |a_{n_k} - a_n| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\]

This proves the theorem.

Definition 16.3.17 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 16.3.18 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,

\[
|x_i - y_i| \leq 2^{-1} |b_i - a_i|.
\]

\[
\text{diam} (I_0) = \left( \sum_{i=1}^{p} |b_i - a_i|^2 \right)^{1/2},
\]

\[
|x - y| = \left( \sum_{i=1}^{p} |x_i - y_i|^2 \right)^{1/2}
\]

\[
\leq 2^{-1} \left( \sum_{i=1}^{p} |b_i - a_i|^2 \right)^{1/2} \equiv 2^{-1} \text{diam} (I_0).
\]

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 \( \{J_1, \ldots, J_{2p}\} \) 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 \( J_1 \). Next do for \( J_1 \) what was done for \( I_0 \) to get \( I_2 \subseteq I_1 \) such that the diameter is half that of \( J_1 \) and \( I_2 \) contains \( \{a_k\} \) for infinitely many values of \( k \). Continue in this way obtaining a nested sequence of intervals, \( \{J_k\} \) 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 16.3.19** Every Cauchy sequence in \( \mathbb{R}^p \) converges.

**Proof:** Let \( \{a_k\} \) be a Cauchy sequence. By Theorem 16.3.15 there is some interval, \( \prod_{i=1}^p [a_i, b_i] \) containing all the terms of \( \{a_k\} \). Therefore, by Theorem 16.3.18 a subsequence converges to a point of this interval. By Theorem 16.3.16 the original sequence converges. This proves the theorem.

**16.3.4 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 16.3.20** A function \( f : D(f) \to \mathbb{R}^q \) 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 \( |x - x_n| < \frac{1}{n} \), 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.
Part V

Differentiability
Chapter 17

Differentiability

17.1 The Definition Of Differentiability

Quiz

1. Let \( f(x, y) = x^2 y + \sin(xy) \). Find \( \nabla f(x, y) \).

2. Let \( f(x, y) = x^2 y + \sin(xy) \). Find \( D_vf(1, 1) \) where \( v \) is in the direction of \((1, 2)\).

3. Let \( f(x, y) = x^2 y + \sin(xy) \). Find the largest value of \( D_vf(1, 2) \) for all \( v \). That is, find the largest directional derivative of this function.

First remember what it means for a function of one variable to be differentiable.

\[
f'(x) \equiv \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}
\]

Another way to say this is contained in the following observation.

**Observation 17.1.1** Suppose a function, \( f \) of one variable has a derivative at \( x \). Then

\[
\lim_{h \to 0} \left| \frac{f(x+h) - f(x) - f'(x)h}{h} \right| = 0.
\]

For a function of \( n \) variables, there is a similar definition of what it means for a function to be differentiable.

**Definition 17.1.2** Let \( U \) be an open set in \( \mathbb{R}^n \) and suppose \( f : U \to \mathbb{R} \) is a function. Then \( f \) is differentiable at \( x \in U \) if for \( v = (v_1, \ldots, v_n) \)

\[
\lim_{|v| \to 0} \frac{\left| f(x + v) - f(x) - \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x) v_k \right|}{|v|} = 0.
\]

Written in terms of matrices, this says

\[
\lim_{|v| \to 0} \left| \frac{f(x + v) - f(x) - \left( \frac{\partial f}{\partial x_1}(x) \cdots \frac{\partial f}{\partial x_n}(x) \right) v}{|v|} \right| = 0.
\]

See how the matrix (vector) plays the same role as \( f'(x) \) for a function of one variable. There you multiply by \( f'(x) \). Here you multiply by the matrix \( \left( \frac{\partial f}{\partial x_1}(x) \cdots \frac{\partial f}{\partial x_n}(x) \right) \). It is often the case, especially when \( f \) has scalar values, that we write it in terms of a dot product.

\[
\left( \frac{\partial f}{\partial x_1}(x) \cdots \frac{\partial f}{\partial x_n}(x) \right) \cdot v
\]
Definition 17.1.3 A function of a vector, \( \mathbf{v} \) is called \( o(\mathbf{v}) \) if
\[
\lim_{|\mathbf{v}| \to 0} \frac{o(\mathbf{v})}{|\mathbf{v}|} = 0.
\] (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
\[
o(\mathbf{v}) = o(\mathbf{v}) + o(\mathbf{v}), o(\mathbf{v}) = 45o(\mathbf{v}), o(\mathbf{v}) = o(\mathbf{v}) - o(\mathbf{v}), etc.
\]

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 obsession over things which are not relevant and should be ignored.

Definition 17.1.2 is then equivalent to the following very simple statement.

Definition 17.1.4 Let \( U \) be an open set in \( \mathbb{R}^n \) and suppose \( f: U \to \mathbb{R} \) is a function. Then \( f \) is differentiable at \( x \in U \) if
\[
f(x + \mathbf{v}) - f(x) - \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x) v_k = o(\mathbf{v}).
\]

The first definition says nothing more than
\[
f(x + \mathbf{v}) - f(x) - \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x) v_k = o(\mathbf{v})
\]
because it says
\[
\lim_{|\mathbf{v}| \to 0} \frac{|f(x + \mathbf{v}) - f(x) - \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x) v_k|}{|\mathbf{v}|} = 0.
\]

The following is fundamental.

Proposition 17.1.5 If \( f \) is differentiable at \( x \), then \( f \) is continuous at \( x \).

Proof: From the definition of differentiability,
\[
|f(x + \mathbf{v}) - f(x)| \leq \sum_{k=1}^{n} \left| \frac{\partial f}{\partial x_k}(x) v_k + o(\mathbf{v}) \right|
\]
Let \( \varepsilon > 0 \) be given. Then clearly if \( |\mathbf{v}| \) is sufficiently small, the right side of the above is less than \( \varepsilon \). Thus the function is continuous at \( x \).

So which functions are differentiable? Are there simple ways to look at a function and say that it is clearly differentiable? Existence of partial derivatives is needed in order to even write the above expression but it turns out this is not enough. Here is a simple example.

Example 17.1.6 Let
\[
f(x, y) = \begin{cases} 
\frac{xy}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0) \\
0 & \text{if } (x, y) = (0, 0)
\end{cases}
\]
Then
\[
f_x(0, 0) = \lim_{h \to 0} \frac{f(h, 0) - f(0, 0)}{h} = \lim_{h \to 0} \frac{0}{h} = 0
\]
Also
\[
f_y(0, 0) = \lim_{h \to 0} \frac{f(0, h) - f(0, 0)}{h} = \lim_{h \to 0} \frac{0}{h} = 0
\]
so both partial derivatives exist. However, the function is not even continuous at \( (0, 0) \). This is because it equals zero on the entire \( y \) axis but along the line, \( y = x \) the function equals \( 1/2 \). By Proposition 17.1.3 it cannot be differentiable.
17.2. $C^1$ FUNCTIONS AND DIFFERENTIABILITY

17.2 $C^1$ Functions And Differentiability

It turns out that if the partial derivatives are continuous then the function is differentiable. I will show this next. First, remember the Cauchy-Schwarz inequality, which I will list here for convenience.

$$\left| \sum_{i=1}^{n} a_i b_i \right| \leq \left( \sum_{i=1}^{n} a_i^2 \right)^{1/2} \left( \sum_{i=1}^{n} b_i^2 \right)^{1/2}.$$

**Theorem 17.2.1** Suppose $f : U \rightarrow \mathbb{R}$ where $U$ is an open set. Suppose also that all partial derivatives of $f$ exist on $U$ and are continuous. Then $f$ is differentiable at every point of $U$.

**Proof:** If you fix all the variables but one, you can apply the fundamental theorem of calculus as follows.

$$f(x+v_k e_k) - f(x) = \int_{0}^{1} \frac{\partial f}{\partial x_k}(x + tv_k e_k) v_k dt. \quad (17.2)$$

Here is why. Let $h(t) = f(x + tv_k e_k)$. Then

$$\frac{h(t + \Delta t) - h(t)}{\Delta t} = \frac{f(x + tv_k e_k + \Delta tv_k e_k) - f(x + tv_k e_k)}{\Delta tv_k} v_k$$

and so, taking the limit as $\Delta t \rightarrow 0$ yields

$$h'(t) = \frac{\partial f}{\partial x_k}(x + tv_k e_k) v_k$$

Therefore,

$$f(x+v_k e_k) - f(x) = h(1) - h(0) = \int_{0}^{1} h'(t) dt = \int_{0}^{1} \frac{\partial f}{\partial x_k}(x + tv_k e_k) v_k dt.$$

Now I will use this observation to prove the theorem. Let $v = (v_1, \ldots, v_n)$ with $|v|$ sufficiently small. Thus $v = \sum_{k=1}^{n} v_k e_k$. For the purposes of this argument, define

$$\sum_{k=n+1}^{n} v_k e_k = 0.$$
Then with this convention, and using (17.2),

\[
\begin{align*}
    f(x + v) - f(x) &= \sum_{i=1}^{n} \left( f \left( x + \sum_{k=i+1}^{n} v_k e_k \right) - f \left( x + \sum_{k=i+1}^{n} v_k e_k \right) \right) \\
    &= \sum_{i=1}^{n} \int_{0}^{1} \frac{\partial f}{\partial x_i} \left( x + \sum_{k=i+1}^{n} v_k e_k + tv_i e_i \right) v_i dt \\
    &= \sum_{i=1}^{n} \int_{0}^{1} \left( \frac{\partial f}{\partial x_i} \left( x + \sum_{k=i+1}^{n} v_k e_k + tv_i e_i \right) v_i - \frac{\partial f}{\partial x_i}(x) v_i \right) dt \\
    &\quad + \sum_{i=1}^{n} \int_{0}^{1} \frac{\partial f}{\partial x_i}(x) v_i dt \\
    &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x) v_i + \int_{0}^{1} \sum_{i=1}^{n} \left( \frac{\partial f}{\partial x_i} \left( x + \sum_{k=i+1}^{n} v_k e_k + tv_i e_i \right) - \frac{\partial f}{\partial x_i}(x) \right) v_i dt \\
    &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x) v_i + o(v)
\end{align*}
\]

and this shows \( f \) is differentiable at \( x \) because it satisfies the conditions of Definition (17.1.4). Some explanation of the step to the last line is in order. The messy thing at the end is \( o(v) \) because of the continuity of the partial derivatives. In fact, from the Cauchy Schwarz inequality,

\[
\begin{align*}
    \int_{0}^{1} \sum_{i=1}^{n} \left( \frac{\partial f}{\partial x_i} \left( x + \sum_{k=i+1}^{n} v_k e_k + tv_i e_i \right) - \frac{\partial f}{\partial x_i}(x) \right) v_i dt \\
    &\leq \int_{0}^{1} \left( \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \left( x + \sum_{k=i+1}^{n} v_k e_k + tv_i e_i \right) - \frac{\partial f}{\partial x_i}(x) \right|^2 \right)^{1/2} dt \left( \sum_{i=1}^{n} v_i^2 \right)^{1/2} \\
    &= \int_{0}^{1} \left( \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \left( x + \sum_{k=i+1}^{n} v_k e_k + tv_i e_i \right) - \frac{\partial f}{\partial x_i}(x) \right|^2 \right)^{1/2} dt |v| \\
    \end{align*}
\]

Thus, dividing by \( |v| \) and taking a limit as \( |v| \to 0 \), the quotient is nothing but

\[
\int_{0}^{1} \left( \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \left( x + \sum_{k=i+1}^{n} v_k e_k + tv_i e_i \right) - \frac{\partial f}{\partial x_i}(x) \right|^2 \right)^{1/2} dt \]

which converges to 0 due to continuity of the partial derivatives of \( f \). This proves the theorem.

To help you keep the various terms straight, here is a pretty diagram.
You might ask whether there are examples of functions which are differentiable but not $C^1$. Of course there are. There are easy examples of this even for functions of one variable. Here is one.

$$\begin{align*}
  f(x) &= \begin{cases} 
    x^2 \sin \frac{1}{x} & \text{if } x \neq 0 \\
    0 & \text{if } x = 0 
  \end{cases} 
\end{align*}$$

You should show that $f'(0) = 0$ but the derivative is $2x \sin \frac{1}{x} - \cos \frac{1}{x}$ for $x \neq 0$ and this function fails to even have a limit as $x \to 0$. This is a great test question. You ask for $f'(0)$ and it is really easy if you use the definition. However, people usually find $f'(x)$ and then try to plug in $x = 0$. This is doomed to failure and makes the question very easy to grade.

### 17.3 The Directional Derivative

Here I will prove the formula for the directional derivative presented earlier. Recall that for $v$ a unit vector, $||v|| = 1$

$$D_v f(x) = \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t}.$$ 

**Theorem 17.3.1** Suppose $f$ is differentiable at $x$. Then $D_v (f) (x)$ exists and is given by

$$D_v (f) (x) = \nabla f(x) \cdot v$$

**Proof:** By differentiability of $f$ at $x$,

$$\frac{f(x + tv) - f(x)}{t} = \frac{1}{t} \left( \sum_{k=1}^{n} \frac{\partial f}{\partial x_k} (x) tv_k + o(tv) \right)$$

$$= \sum_{k=1}^{n} \frac{\partial f}{\partial x_k} (x) v_k + \frac{o(tv)}{|tv|}$$

Taking the limit as $t \to 0$,

$$D_v f(x) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k} (x) v_k \equiv \nabla f(x) \cdot v.$$

This proves the theorem.

What is the direction in which the largest directional derivative results? You want to maximize $\nabla f(x) \cdot v = |\nabla f(x)| |v| \cos \theta$ where $\theta$ is the included angle between $v$ and $\nabla f(x)$.
Clearly this occurs when \( \theta = 0 \). Therefore, the largest value of the directional derivative is when \( v = \nabla f (x) / |\nabla f (x)| \). The value of the directional derivative in this direction, is

\[
\nabla f (x) \cdot \nabla f (x) / |\nabla f (x)| = |\nabla f (x)|.
\]

Similarly, the smallest value for the directional derivative occurs when \( v = -\nabla f (x) / |\nabla f (x)| \) because this corresponds to \( \theta = \pi \) and \( \cos \theta = -1 \). The smallest value of the directional derivative is then \(-|\nabla f (x)|\).

### 17.3.1 Separable Differential Equations

If you do not know how to solve simple differential equations, read this section. Otherwise skip it.

Differential equations are just equations which involve an unknown function and some of its derivatives. For example, a differential equation is

\[
y' = 1 + y^2.
\]

You might check and see that a solution to this differential equation is \( y = \tan x \).

Here is another easier differential equation.

\[
y' = x^2.
\]

A solution to this one is of the form

\[
y = \frac{x^3}{3} + C
\]

where \( C \) is any constant. In general, you are familiar with differential equations of the form

\[
y' = f (x).
\]

The problem is just to find an antiderivative of the given function. However, equations like the one in (17.3) are not so obvious. It turns out there are many recipes for finding solutions to differential equations of various sorts. One of the easiest kinds of differential equations to solve are those which are separable.

Separable differential equations are those which can be written in the form

\[
\frac{dy}{dx} = \frac{f (x)}{g (y)}.
\]

The equation in (17.3) is an example of a separable differential equation. Just let \( f (x) = 1 \) and \( g (y) = \frac{1}{1+y^2} \).

The reason these are called separable is that if you formally cross multiply,

\[
g (y) dy = f (x) dx
\]

and the variables are “separated”. Here is how you solve these. Find \( G'(y) = g (y) \) and \( F'(x) = f (x) \). That is, pick \( G \in \int g (y) dy \) and \( F \in \int f (x) dx \). Suppose \( F (x) - G (y) = c \) specifies \( y \) as a differentiable function of \( x \), then \( x \rightarrow y (x) \) solves the separable differential equation because by the chain rule,

\[
F'(x) - G'(y) y' = f (x) - g (y) y'
\]

and so

\[
f (x) = g (y) y'
\]
so
\[ y' = \frac{f(x)}{g(y)}. \]

This is why the solutions are given in the form
\[ F(x) - G(y) = c \]
where \( c \) is a constant, or equivalently
\[ G(y) - F(x) = c \]
where \( c \) is a constant.

**Example 17.3.2** Find the solutions to the differential equation,
\[ y' = \frac{x^2}{y} \]
which satisfies the initial condition, \( y(0) = 1 \). Since there is a differential equation along with an initial condition, this is called an initial value problem.

To solve this you separate the variables and write
\[ ydy = x^2dx \]
and then from the above discussion,
\[ \frac{y^2}{2} - \frac{x^3}{3} = C. \tag{17.4} \]
You want \( y = 1 \) when \( x = 0 \) and so you must have
\[ \frac{1}{2} = C. \]
The solution is
\[ y = \sqrt{2 \left( \frac{x^3}{3} + \frac{1}{2} \right)} \]
where it was necessary to pick the positive square root because otherwise, you would not have \( y(0) = 1 \).

Sometimes you can’t solve for \( y \) in terms of \( x \).

**Example 17.3.3** Find the solutions to the differential equation,
\[ y' = \frac{x^2}{y \sin y}. \]

In this case,
\[ (y \sin y) \, dy = x^2 \, dx \]
and so \( \int y \sin y = \sin y - y \cos y \)
\[ \sin y - y \cos y - \frac{x^3}{3} = C \]
gives the solutions. I would not like to try and solve this for \( y \) in terms of \( x \). Therefore, in this case, it is customary to leave the solution in this form and refer to it as an implicitly defined solution. The point is, the above equation does define \( y \) as a function of \( x \) near typical points on the level curve but it might not be possible to algebraically find \( y \) as a function of \( x \). You notice in the above argument for finding solutions, it was never assumed that you could algebraically find \( y \) as a function of \( x \) in \( F(x) - G(y) = C \).

Here is an interesting example which is non trivial.
Example 17.3.4 What is the equation of a hanging chain?

Consider the following picture of a portion of this chain.

In this picture, \( \rho \) denotes the density of the chain which is assumed to be constant and \( g \) is the acceleration due to gravity. \( T(x) \) and \( T_0 \) represent the magnitude of the tension in the chain at \( x \) and at 0 respectively, as shown. Let the bottom of the chain be at the origin as shown. If this chain does not move, then all these forces acting on it must balance. In particular,

\[
T(x) \sin \theta = l(x) \rho g, \quad T(x) \cos \theta = T_0.
\]

Therefore, dividing these yields

\[
\frac{\sin \theta}{\cos \theta} = l(x) \frac{\rho g}{T_0}.
\]

Now letting \( y(x) \) denote the \( y \) coordinate of the hanging chain corresponding to \( x \),

\[
\frac{\sin \theta}{\cos \theta} = \tan \theta = y'(x).
\]

Therefore, this yields

\[
y'(x) = cl(x).
\]

Now differentiating both sides of the differential equation,

\[
y''(x) = cl'(x) = c\sqrt{1 + y'(x)^2}
\]

and so

\[
\frac{y''(x)}{\sqrt{1 + y'(x)^2}} = c.
\]

Let \( z(x) = y'(x) \) so the above differential equation becomes

\[
\frac{dz}{dx} = c\sqrt{1 + z^2},
\]

a separable differential equation. Thus

\[
\frac{dz}{\sqrt{1 + z^2}} = cdx.
\]

Now \( \int \frac{dz}{\sqrt{1 + z^2}} = \text{arcsinh}(z) + C \) and so the solutions are of the form

\[
\text{arcsinh}(z) - cx = d
\]
where \(d\) is some constant. Thus

\[ y' = z = \sinh (cx + d) \]

and so

\[ y(x) \in \int \sinh (cx + C) \, dx = \frac{\cosh (cx + d)}{c} + k \]

where \(k\) is some constant. Therefore,

\[ y(x) = \frac{1}{c} \cosh (cx + d) + k \]

where \(d\) and \(k\) are some constants and \(c = \rho g / T_0\). Curves of this sort are called catenaries. Note these curves result from an assumption the only forces acting on the chain are as shown.

### 17.4 Exercises With Answers*

1. Find the solution to the initial value problem,

\[ y' = \frac{x}{y^2}, \quad y(0) = 1. \]

Separating the variables, you get \(y^2 \, dy = x \, dx\) and so \(\frac{y^3}{3} - \frac{x^2}{2} = c\). From the initial condition, \(\frac{1}{3} = c\) and so the solution is

\[ \frac{y^3}{3} - \frac{x^2}{2} = \frac{1}{3} \]

2. Find the solution to the initial value problem,

\[ \tan (y) \, y' = \sin x, \quad y \left( \frac{\pi}{4} \right) = \frac{\pi}{4}. \]

Separating the variables, \(\tan (y) \, dy = \sin (x) \, dx\) and so \(\ln |\sec (y)| + \cos (x) = c\). Now from the initial condition,

\[ \ln \left( \sqrt{2} \right) + \frac{\sqrt{2}}{2} = c \]

and so \(\ln |\sec (y)| + \cos (x) = \ln \left( \sqrt{2} \right) + \frac{\sqrt{2}}{2} \)

3. Find the solution to the initial value problem,

\[ y' = \frac{y}{x}, \quad y(1) = 1. \]

Separating the variables, gives \(\frac{dy}{y} = \frac{dx}{x}\) and so \(y - x = c\). But from the initial condition, \(c = 0\). Hence \(y = x\).
17.4.1 A Heat Seaking Particle

Suppose the temperature is given as $T(x,y,z)$ and a particle tries to go in the direction of most rapid rate of change of temperature. In other words this particle likes it hot. This means it moves in the direction of the gradient of $T$. In other words, $(x', y', z')^T = k(x,y,z) \nabla T(x,y,z)$.

Of course you don’t know what $k(x,y,z)$ is but if you did and if you also knew $T$, then you would have a system of differential equations for the position of the particle as a function of time. If you were given an initial position, you could then ask for the solution to the resulting intial value problem. Of course you won’t be able to solve the equations in general. These sorts of things require numerical methods. Also, in interesting examples, everything would also depend on $t$. The following pseudo application has to do with a situation which I will cook up so that I will be able to solve everything.

Example 17.4.1 A heat seaking particle starts at $(1, 2, 1)$. The temperature is $T(x,y,z) = x^2 + y + z^3$ and assume that $k = 1$. Find the motion of the heat seaking particle.

As explained above, you need $(x', y', z')^T = \nabla T(x,y,z)$ and so

\[
\frac{dx}{dt} = 2x, \quad \frac{dy}{dt} = 1, \quad \frac{dz}{dt} = 3z^2
\]

because $\nabla T = (2x, 1, 3z^2)^T$. It is very fortunate that the equations are not coupled. Consider the first one. Separating the variables,

\[
\frac{dx}{x} = 2dt
\]

and so $\ln(x) - 2t = c$. From the initial condition which states that at $t = 0, x = 1$, it follows $c = 0$. Therefore, $x = e^{2t}$. Next consider the second of the differential equations. This one says $y = t + c$ and from the initial condition, $c = 2$ so the second gives $y = t + 2$. Finally the last equation separates to give

\[
\frac{dz}{z^2} = 3dt
\]

and so

\[
-\frac{1}{z} = 3t + c.
\]

In this case the initial data gives $c = -1$. Therefore, $z = -\frac{1}{3t-1}$. It follows the path of the particle is of the form

\[
\left(e^{2t}, t + 2, -\frac{1}{3t-1}\right).
\]

Note that this only makes sense for $t \in \left[0, \frac{1}{3}\right)$. This type of thing is typical of nonlinear differential equations.

I think you can see how to do similar problems in which the particle is heat avoiding. You just put in a minus sign by $\nabla T$.

17.5 The Chain Rule

Remember what this was all about for a function of one variable. You had $z = f(y)$ and $y = g(x)$ and you wanted to find $\frac{dz}{dx}$. Remember the answer was

\[
\frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{dx}.
\]
The chain rule was one of the most important rules for differentiation. Its importance is no less for functions of many variables.

The problem is this: \( z = f(y_1, y_2, \ldots, y_n) \) and \( y_k = g_k(x_1, \ldots, x_m) \). You want to find \( \frac{\partial z}{\partial x_k} \) for each \( k = 1, 2, \ldots, m \). It turns out to be exactly the same sort of formula which works. In this case the formula is

\[
\frac{\partial z}{\partial x_k} = \sum_{i=1}^{n} \frac{\partial z}{\partial y_i} \frac{\partial y_i}{\partial x_k}.
\]

People who use the repeated index summation convention write this as

\[
\frac{\partial z}{\partial x_k} = \frac{\partial z}{\partial y_i} \frac{\partial y_i}{\partial x_k}
\]

which is formally just like it was for a function of one variable. I think this is one reason for the attractiveness of this repeated summation convention. Here is an example.

**Example 17.5.1** Suppose \( z = y_1 + y_2y_3^2 \) and \( y_1 = \sin(x_1) + x_2, y_2 = \cos(x_3), \) and \( y_3 = x_1^2 + \sin x_2 + x_4 \). Find \( \frac{\partial z}{\partial x_2} \) and \( \frac{\partial z}{\partial x_4} \).

From the above formula,

\[
\frac{\partial z}{\partial x_2} = \frac{\partial z}{\partial y_1} \frac{\partial y_1}{\partial x_2} + \frac{\partial z}{\partial y_2} \frac{\partial y_2}{\partial x_2} + \frac{\partial z}{\partial y_3} \frac{\partial y_3}{\partial x_2}
\]

\[
= 1 \times 1 + y_3^2 \times 0 + 2y_2y_3 \cos x_2
\]

\[
= 1 + 2y_2y_3 \cos x_2.
\]

If you want to put this in terms of the \( x \) variables, it is

\[
\frac{\partial z}{\partial x_2} = 1 + 2y_2y_3 \cos x_2
\]

\[
= 1 + 2 \cos(x_3) (x_1^2 + \sin x_2 + x_4) \cos x_2.
\]

Now consider the other partial derivative.

\[
\frac{\partial z}{\partial x_4} = \frac{\partial z}{\partial y_1} \frac{\partial y_1}{\partial x_4} + \frac{\partial z}{\partial y_2} \frac{\partial y_2}{\partial x_4} + \frac{\partial z}{\partial y_3} \frac{\partial y_3}{\partial x_4}
\]

\[
= \frac{\partial z}{\partial y_1} \times 0 + \frac{\partial z}{\partial y_2} \times 0 + 2y_2y_3 \times 1
\]

\[
= 2 \cos(x_3) (x_1^2 + \sin x_2 + x_4) .
\]

Be sure you can find and place the partial derivatives in terms of the independent variables, \( x \). It is just as correct to leave the answer in terms of \( y \) and \( x \) but sometimes people may insist you place the answer in terms of \( x \).

The next task is to explain why the above formula works. The argument I will give applies to one dimension also. Therefore, you can consider it a review of what you should have seen in beginning calculus.

**Lemma 17.5.2** Suppose \( U \) is an open set in \( \mathbb{R}^n \) and \( f : U \to \mathbb{R} \). Suppose \( x \in U \) and for all \( v \) small enough,

\[
f(x + v) - f(x) = \sum_{i=1}^{n} a_i v_i + o(v).
\]

Then \( a_i = \frac{\partial f}{\partial x_i}(x) \) and \( f \) is differentiable.
CHAPTER 17. DIFFERENTIABILITY

Proof: Let \( t \) be a small nonzero number. Then since the \( i^{th} \) component of \( e_k \) equals zero unless \( i = k \) when it is 1,
\[
    f(x + te_k) - f(x) = a_k t + o(te_k) = a_k t + o(t)
\]
Now divide by \( t \) and take a limit.
\[
    \frac{\partial f}{\partial x_k}(x) = \lim_{t \to 0} \frac{f(x + te_k) - f(x)}{t} = \lim_{t \to 0} \left( a_k + \frac{o(t)}{t} \right) = a_k.
\]
This proves the lemma.

Lemma 17.5.3 Let \( U \) be an open set and suppose \( g \) is differentiable at \( x \in U \). Then
\[
    o(g(x + v) - g(x)) = o(v).
\]
Proof: I need to show
\[
    \lim_{|v| \to 0} \frac{o(g(x + v) - g(x))}{|v|} = 0.
\]
Let \( \varepsilon > 0 \) be given. Since \( g \) is continuous at \( x \), there exists \( \delta_1 > 0 \) such that if \( |v| < \delta_1 \), then
\[
    \frac{|o(g(x + v) - g(x))|}{|g(x + v) - g(x)|} < \varepsilon
\]
Hence, for such \( v \),
\[
    |o(g(x + v) - g(x))| < \varepsilon |g(x + v) - g(x)|
\]
Since \( g \) is differentiable at \( x \), there exists \( \delta_2 > 0 \) such that if \( |v| < \delta_2 \),
\[
    \left| \frac{g(x + v) - g(x) - \sum_{k=1}^{n} \frac{\partial g}{\partial x_k}(x) v_k}{|v|} \right| < \varepsilon
\]
Hence for \( |v| < \delta_2 \),
\[
    |g(x + v) - g(x)| < \sum_{k=1}^{n} \left| \frac{\partial g}{\partial x_k}(x) v_k \right| + \varepsilon |v| \quad (17.6)
\]
Let \( \delta \leq \min(\delta_1, \delta_2) \). Then if \( |v| < \delta \), both (17.5) and (17.6) hold and so by the Cauchy Schwarz inequality,
\[
    |o(g(x + v) - g(x))| < \varepsilon |g(x + v) - g(x)|
\]
\[
    < \varepsilon \left( \sum_{k=1}^{n} \left| \frac{\partial g}{\partial x_k}(x) v_k \right| + \varepsilon |v| \right)
\]
\[
    \leq \varepsilon \left( \sum_{k=1}^{n} \left| \frac{\partial g}{\partial x_k}(x) \right|^2 \right)^{1/2} |v| + \varepsilon |v|.
\]
Dividing both sides by \( |v| \),
\[
    \frac{o(g(x + v) - g(x))}{|v|} \leq \varepsilon \left( \sum_{k=1}^{n} \left| \frac{\partial g}{\partial x_k}(x) \right|^2 \right)^{1/2} + 1
\]
and since \( \varepsilon > 0 \) is arbitrary, this establishes the lemma.

With these lemmas, it is easy to prove the chain rule.
Theorem 17.5.4 Let $V$ be an open set in $\mathbb{R}^n$ and let $U$ be an open set in $\mathbb{R}^m$. Also let $g: U \to V$ be a vector valued function having the property that for each $g_k$ is differentiable at $x \in U$. Also suppose $f: V \to \mathbb{R}$ is differentiable at $g(x)$. Then for $z \equiv f \circ g$, $y_i = g_i(x)$,

$$\frac{\partial z}{\partial x_k}(x) = \sum_{i=1}^{n} \frac{\partial z}{\partial y_i}(g(x)) \frac{\partial y_i}{\partial x_k}(x).$$

Proof: Using Lemma 17.5.2 as needed,

$$f \circ g(x + v) - f \circ g(x) = \sum_{i=1}^{n} \frac{\partial z}{\partial y_i}(g(x))(g_i(x + v) - g_i(x)) + o(g_i(x + v) - g_i(x))$$

$$= \sum_{i=1}^{n} \frac{\partial z}{\partial y_i}(g(x))(g_i(x + v) - g_i(x)) + o(v)$$

$$= \sum_{i=1}^{n} \frac{\partial z}{\partial y_i}(g(x)) \left( \sum_{k=1}^{m} \frac{\partial y_i}{\partial x_k}(x) v_k + o(v) \right) + o(v)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{m} \frac{\partial z}{\partial y_i}(g(x)) \frac{\partial y_i}{\partial x_k}(x) v_k + o(v)$$

Now by Lemma 17.5.2,

$$\frac{\partial f \circ g}{\partial x_k}(x) = \sum_{i=1}^{n} \frac{\partial z}{\partial y_i}(g(x)) \frac{\partial y_i}{\partial x_k}(x).$$

This proves the theorem.

17.5.1 Related Rates Problems

Sometimes several variables are related and given information about how one variable is changing, you want to find how the others are changing. The following law is discussed later in the book, on Page 539.

Example 17.5.5 Bernoulli’s law states that in an incompressible fluid,

$$\frac{v^2}{2g} + z + \frac{P}{\gamma} = C$$

where $C$ is a constant. Here $v$ is the speed, $P$ is the pressure, and $z$ is the height above some reference point. The constants, $g$ and $\gamma$ are the acceleration of gravity and the weight density of the fluid. Suppose measurements indicate that $\frac{dv}{dt} = -3$, and $\frac{dz}{dt} = 2$. Find $\frac{dP}{dt}$ when $v = 7$ and $z = 8$ in terms of $g$ and $\gamma$.

This is just an exercise in using the chain rule. Differentiate the two sides with respect to $t$.

$$\frac{1}{g} \frac{dv}{dt} + \frac{dz}{dt} + \frac{1}{\gamma} \frac{dP}{dt} = 0.$$
Then when \( v = 7 \) and \( z = 8 \), finding \( \frac{dP}{dt} \) involves nothing more than solving the following for \( \frac{dP}{dt} \).

\[
\frac{7}{g} (-3) + 2 + \frac{1}{\gamma} \frac{dP}{dt} = 0
\]

Thus

\[
\frac{dP}{dt} = \gamma \left( \frac{21}{g} - 2 \right)
\]

at this instant in time.

**Example 17.5.6** In Bernoulli’s law above, each of \( v, z, \) and \( P \) are functions of \((x, y, z)\), the position of a point in the fluid. Find a formula for \( \frac{\partial P}{\partial x} \) in terms of the partial derivatives of the other variables.

This is an example of the chain rule. Differentiate both sides with respect to \( x \).

\[
\frac{v}{g} v_x + z_x + \frac{1}{\gamma} P_x = 0
\]

and so

\[
P_x = -\left( \frac{vv_x + zz_x g}{g} \right) \gamma
\]

**Example 17.5.7** Suppose a level curve is of the form \( f(x, y) = C \) and that near a point on this level curve, \( y \) is a differentiable function of \( x \). Find \( \frac{dy}{dx} \).

This is an example of the chain rule. Differentiate both sides with respect to \( x \). This gives

\[
f_x + f_y \frac{dy}{dx} = 0.
\]

Solving for \( \frac{dy}{dx} \) gives

\[
\frac{dy}{dx} = -\frac{f_x (x, y)}{f_y (x, y)}.
\]

**Example 17.5.8** Suppose a level surface is of the form \( f(x, y, z) = C \) and that near a point, \((x, y, z)\) on this level surface, \( z \) is a \( C^1 \) function of \( x \) and \( y \). Find a formula for \( z_x \).

This is an example of the use of the chain rule. Differentiate both sides of the equation with respect to \( x \). Since \( y_x = 0 \), this yields

\[
f_x + f_z z_x = 0.
\]

Then solving for \( z_x \) gives

\[
z_x = -\frac{f_x (x, y, z)}{f_z (x, y, z)}
\]

**Example 17.5.9** Polar coordinates are

\[
x = r \cos \theta, \ y = r \sin \theta.
\]

Thus if \( f \) is a \( C^1 \) scalar valued function you could ask to express \( f_x \) in terms of the variables, \( r \) and \( \theta \). Do so.
This is an example of the chain rule. \( f = f(r, \theta) \) and so

\[
f_x = f_r r_x + f_\theta \theta_x.
\]

This will be done if you can find \( r_x \) and \( \theta_x \). However you must find these in terms of \( r \) and \( \theta \), not in terms of \( x \) and \( y \).

Using the chain rule on the two equations for the transformation,

\[
1 = r_x \cos \theta - (r \sin \theta) \theta_x
\]

\[
0 = r_x \sin \theta + (r \cos \theta) \theta_x
\]

Solving these using Cramer’s rule yields

\[
r_x = \cos(\theta), \quad \theta_x = \frac{-\sin(\theta)}{r}
\]

Hence \( f_x \) in polar coordinates is

\[
f_x = f_r(r, \theta) \cos(\theta) - f_\theta(r, \theta) \left( \frac{\sin(\theta)}{r} \right)
\]

### 17.6 Taylor’s Formula An Application Of Chain Rule

Now recall the Taylor formula with the Lagrange form of the remainder. What follows is a proof of this important result based on the mean value theorem or Rolle’s theorem.

**Theorem 17.6.1** Suppose \( f \) has \( n + 1 \) derivatives on an interval, \((a,b)\) and let \( c \in (a,b) \). Then if \( x \in (a,b) \), there exists \( \xi \) between \( c \) and \( x \) such that

\[
f(x) = f(c) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!} (x-c)^k + f^{(n+1)}(\xi) \frac{(x-c)^{n+1}}{(n+1)!}.
\]

(In this formula, the symbol \( \sum_{k=1}^{0} a_k \) will denote the number 0.)

**Proof:** It can be assumed \( x \neq c \) because if \( x = c \) there is nothing to show. Then there exists \( K \) such that

\[
f(x) - 
\left( f(c) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!} (x-c)^k + K(x-c)^{n+1} \right) = 0
\]

(17.7)

In fact,

\[
K = -\frac{f(x) - \left( f(c) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!} (x-c)^k \right)}{(x-c)^{n+1}}.
\]

Now define \( F(t) \) for \( t \) in the closed interval determined by \( x \) and \( c \) by

\[
F(t) \equiv f(x) - \left( f(t) + \sum_{k=1}^{n} \frac{f^{(k)}(t)}{k!} (x-t)^k + K(x-t)^{n+1} \right).
\]

The \( c \) in \( \text{[17.7]} \) got replaced by \( t \).
Therefore, \( F(c) = 0 \) by the way \( K \) was chosen and also \( F(x) = 0 \). By the mean value theorem or Rolle’s theorem, there exists \( \xi \) between \( x \) and \( c \) such that \( F'(\xi) = 0 \). Therefore,

\[
0 = f'(\xi) + \sum_{k=1}^{n} f^{(k+1)}(\xi) \frac{(x-\xi)^k}{k!} - \sum_{k=1}^{n} f^{(k)}(\xi) \frac{(x-\xi)^k}{(k-1)!} - K(n+1)(x-\xi)^n
\]

\[
= f'(\xi) + \sum_{k=1}^{n} f^{(k+1)}(\xi) \frac{(x-\xi)^k}{k!} - \sum_{k=0}^{n-1} f^{(k+1)}(\xi) \frac{(x-\xi)^k}{k!} - K(n+1)(x-\xi)^n
\]

\[
= f'(\xi) + \frac{f^{(n+1)}(\xi)}{n!} (x-\xi)^n - f'(\xi) - K(n+1)(x-\xi)^n
\]

\[
= \frac{f^{(n+1)}(\xi)}{n!} (x-\xi)^n - K(n+1)(x-\xi)^n
\]

Then therefore,

\[
K = \frac{f^{(n+1)}(\xi)}{(n+1)!}
\]

This proves the theorem.

The following is a special case and is what will be used.

**Theorem 17.6.2** Let \( h : (-\delta, 1+\delta) \rightarrow \mathbb{R} \) have \( m+1 \) derivatives. Then there exists \( t \in (0,1) \) such that

\[
h(1) = h(0) + \sum_{k=1}^{m} \frac{h^{(k)}(0)}{k!} + \frac{h^{(m+1)}(t)}{(m+1)!}.
\]

Now let \( f : U \rightarrow \mathbb{R} \) where \( U \subseteq \mathbb{R}^n \) and suppose \( f \in C^m(U) \). Let \( x \in U \) and let \( r > 0 \) be such that

\[
B(x,r) \subseteq U.
\]

Then for \( |v| < r \), consider

\[
f(x+tv) - f(x) = h(t)
\]

for \( t \in [0,1] \). Then by the chain rule,

\[
h'(t) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x+tv) v_k
\]

\[
h''(t) = \sum_{k_1=1}^{n} \sum_{k_2=1}^{n} \frac{\partial^2 f}{\partial x_{k_1} \partial x_{k_2}}(x+tv) v_{k_1} v_{k_2}
\]

\[
h'''(t) = \sum_{k_1=1}^{n} \sum_{k_2=1}^{n} \sum_{k_3=1}^{n} \frac{\partial^3 f}{\partial x_{k_1} \partial x_{k_2} \partial x_{k_3}}(x+tv) v_{k_1} v_{k_2} v_{k_3}
\]

and in general,

\[
h^{(m)}(t) = \sum_{k_1, k_2, \ldots, k_m} f_{x_{k_2} \cdots x_{k_m}}(x+tv) v_{k_1} v_{k_2} \cdots v_{k_m}
\]

Then with this information, the following is Taylor’s formula for a function of \( n \) variables.
Theorem 17.6.3 Let \( f : U \to \mathbb{R} \) and let \( f \in C^{m+1}(U) \). Then if
\[
B(x, r) \subseteq U,
\]
and \(|v| < r\), there exists \( t \in (0, 1) \) such that.
\[
f(x + v) = f(x) + \sum_{k=0}^{n} \frac{\partial f}{\partial x_k}(x) v_k + \frac{1}{2!} \sum_{k=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_k}(x) v_k v_j \\
+ \frac{1}{3!} \sum_{k_1=1}^{n} \sum_{k_2=1}^{n} \sum_{k_3=1}^{n} \frac{\partial^3 f}{\partial x_{k_1} \partial x_{k_2} \partial x_{k_3}}(x) v_{k_1} v_{k_2} v_{k_3} + \cdots \\
+ \frac{1}{m!} \sum_{k_1, k_2, \ldots, k_m} f_{x_{k_1} x_{k_2} \cdots x_{k_m}}(x) v_{k_1} v_{k_2} \cdots v_{k_m} + \\
\frac{1}{(m+1)!} \sum_{k_1, k_2, \ldots, k_m, k_{m+1}} f_{x_{k_1} x_{k_2} \cdots x_{k_{m+1}}} (x + tv) v_{k_{m+1}} v_{k_1} v_{k_2} \cdots v_{k_m}.
\]
The formulas are long and ugly so people sometimes shorten them as follows.
\[
\sum_{k_1, k_2, \ldots, k_m} f_{x_{k_1} x_{k_2} \cdots x_{k_m}}(x) v_{k_1} v_{k_2} \cdots v_{k_m} \equiv 
\]
\[
D^m f(x) (v, v, \ldots, v)
\]
or to be more brutal,
\[
D^m f(x) v^m
\]
Then the Taylor formula looks better written as
\[
f(x + v) = f(x) + D^1 f(x) v + \frac{1}{2!} D^2 f(x) v^2 + \\
\cdots + \frac{1}{m!} D^m f(x) v^m + \frac{1}{(m+1)!} D^{m+1} f(x + tv) v^{m+1}
\]
for some \( t \in (0, 1) \).

The case where the expansion is stopped with second derivatives is the most useful. In this case you get
\[
f(x + v) = f(x) + D^1 f(x) v + \frac{1}{2!} D^2 f(x + tv) v^2
\]
and that last term is of the form
\[
\frac{1}{2} \begin{pmatrix} v_1 & v_2 \end{pmatrix} \begin{pmatrix} f_{xx}(x + tv) & f_{xy}(x + tv) \\ f_{yx}(x + tv) & f_{yy}(x + tv) \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}
\]
This \( 2 \times 2 \) matrix is called the Hessian matrix. We always assume the mixed partial derivatives are continuous and so it would be unchanged if \( f_{yx} \) were replaced with \( f_{xy} \). This Hessian matrix is important in the second derivative test below.
17.7 Normal Vectors And Tangent Planes

Quiz

1. Let \( z = xy^2 \) and let \( x = ts + ps \) while \( y = 2s^2 + t \). Find \( \frac{\partial z}{\partial t} \) when \( (s, t, p) = (1, 1, 1) \).

2. A level surface is given by \( x^3y + z^2 = 2 \). Find \( z_x \) at the point \((1, 1, 1)\) on the level surface.

3. Suppose \( x = t^3 + s \) and \( y = s^3 + t \). Find \( \frac{\partial z}{\partial t} \) completely in terms of partial derivatives and functions of the new variables, \( s, t \).

The gradient has fundamental geometric significance illustrated by the following picture.

\[
\nabla f(x_0, y_0, z_0)
\]

\[
x_1'(t_0)
\]

\[
x_2'(s_0)
\]

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 surface⁴. 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

\[
f(x_1(t), y_1(t), z_1(t)) = c
\]

for all \( t \) in some interval. Therefore, taking the derivative of both sides and using the chain rule on the left,

\[
\frac{\partial f}{\partial x}(x_1(t), y_1(t), z_1(t))x'_1(t) + \frac{\partial f}{\partial y}(x_1(t), y_1(t), z_1(t))y'_1(t) + \frac{\partial f}{\partial z}(x_1(t), y_1(t), z_1(t))z'_1(t) = 0.
\]

In terms of the gradient, this merely states

\[
\nabla f(x_1(t), y_1(t), z_1(t)) \cdot x'_1(t) = 0.
\]

Similarly,

\[
\nabla f(x_2(s), y_2(s), z_2(s)) \cdot x'_2(s) = 0.
\]

Letting \( s = s_0 \) and \( t = t_0 \), it follows

\[
\nabla f(x_0, y_0, z_0) \cdot x'_1(t_0) = 0, \quad \nabla f(x_0, y_0, z_0) \cdot x'_2(s_0) = 0.
\]

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

---

⁴Do there exist any smooth curves which lie in the level surface of \( f \) and pass through the point \((x_0, y_0, z_0)\)? It turns out there do if \( \nabla f(x_0, y_0, z_0) \neq 0 \) and if the function, \( f \), is \( C^1 \). However, this is a consequence of the implicit function theorem, one of the greatest theorems in all mathematics and a topic for an advanced calculus class. See the the section on the implicit function theorem for the most elementary treatment of this theorem that I know.
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 17.7.1** 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 17.7.2** 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 obtuse angle. The normals are \((2\sqrt{3}, 2, -1)\) and \((2\sqrt{3}, 2, 1)\). Therefore, the cosine of the angle desired is

\[
\frac{(2\sqrt{3})^2 + 4 - 1}{17} = \frac{15}{17}.
\]

**Example 17.7.3** 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 that is needed 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 desired line is

\[
\left(1, \sqrt{3}, 4\right) + t \left(2, 2\sqrt{3}, -1\right).
\]
Chapter 18

Extrema Of Functions Of Several Variables

Quiz

1. Let \( z = x^2 \sin(y) \) and let \( x = t^2 s + r \) while \( y = t^2 - s \). Find \( z_t \) when \((s, t, r) = (1, 1, 1)\).

2. The ideal gas law is \( PV = kT \) where \( k \) is a constant. Suppose at some time \( \frac{dT}{dt} = 1 \), \( \frac{dP}{dt} = -1 \). Find \( \frac{dV}{dt} \) at this instant if \( P = 2, V = 6, T = 100 \).

3. There are two surfaces, \( x^2 + y^2 = 1 \) and \( x^2 + y^2 + z^2 = 5 \) which intersect in a curve. Find an equation of the tangent line to this curve at the point, \( \left( \sqrt{\frac{3}{2}}, \frac{1}{2}, 2 \right) \).

4. A level surface is \( x^2 + 2y^2 + 3z^2 = 6 \). Find the tangent plane at the point, \((1, 1, 1)\).

5. Let \( z = x \sin(x^2 + y^2) \). Find \( \frac{\partial z}{\partial x} \).

6. Let \( z^3 \sin(x) + y^4 z = 7 \). Find \( \frac{\partial z}{\partial x} \).

7. Suppose \( f(x, y) \) is given by
\[
f(x, y) = \begin{cases} 
\frac{2xy + x^2 + xy^2}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0) \\
0 & \text{if } (x, y) = (0, 0)
\end{cases}
\]
Find \( f_x(0, 0) \) if possible.

8. Find parametric equations of the line which is perpendicular to the surface \( 3x^2 + 2y^2 + z^2 = 6 \) at the point \((1, 1, 1)\).

9. Let \( u(x, y) = f(x + y) + g(x - y) \). Compute \( u_{xx} - u_{yy} \). D’Lambert did this problem back in the mid 1700’s and it turned out to be very important.

10. A function of two variables, \( f(x, y) \) is called homogeneous of degree \( \alpha \) if \( f(tx, ty) = t^\alpha f(x, y) \). Establish Euler’s identity which states that for such homogeneous functions,
\[
x f_x(x, y) + y f_y(x, y) = \alpha f(x, y).
\]
This identity dates from early in the 1700’s also. **Hint:** Use the chain rule and differentiate both sides of \( f(tx, ty) = t^\alpha f(x, y) \) with respect to \( t \) using the chain rule and then plug in \( t = 1 \).

Suppose \( f : D(f) \to \mathbb{R} \) where \( D(f) \subseteq \mathbb{R}^n \).
18.1 Local Extrema

Definition 18.1.1 A point \( \mathbf{x} \in D(f) \subseteq \mathbb{R}^n \) is called a **local minimum** if \( f(\mathbf{x}) \leq f(\mathbf{y}) \) for all \( \mathbf{y} \in D(f) \) sufficiently close to \( \mathbf{x} \). A point \( \mathbf{x} \in D(f) \) is called a **local maximum** if \( f(\mathbf{x}) \geq f(\mathbf{y}) \) for all \( \mathbf{y} \in D(f) \) sufficiently close to \( \mathbf{x} \). A **local extremum** is a point of \( D(f) \) which is either a local minimum or a local maximum. The plural for extremum is **minima** and the plural for maximum is **maxima**.

Procedure 18.1.2 To find candidates for local extrema which are interior points of \( D(f) \) where \( f \) is a differentiable 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 \( F(x,z) \equiv z - f(x) = 0 \) 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 \( \mathbf{x} \) is such a point,

\[
k(0, \cdots, 0, 1) = (-f_{x_1}(\mathbf{x}), \cdots, -f_{x_n}(\mathbf{x}), 1).
\]

Thus \( \nabla f(\mathbf{x}) = 0 \).

This is illustrated in the following picture.

A more rigorous explanation is as follows. Let \( \mathbf{v} \) be any vector in \( \mathbb{R}^n \) and suppose \( \mathbf{x} \) is a local maximum (minimum) for \( f \). Then consider the real valued function of one variable, \( h(t) = f(\mathbf{x} + t\mathbf{v}) \) for small \( |t| \). Since \( f \) has a local maximum (minimum), it follows that \( h \) is a differentiable function of the single variable \( t \) for small \( t \) which has a local maximum (minimum) when \( t = 0 \). Therefore, \( h'(0) = 0 \). But \( h'(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{x} + t\mathbf{v}) \mathbf{v}_i \) by the chain rule. Therefore,

\[
h'(0) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{x}) \mathbf{v}_i = 0
\]

and since \( \mathbf{v} \) is arbitrary, it follows \( \frac{\partial f}{\partial x_i}(\mathbf{x}) = 0 \) for each \( i \). Thus

\[
\left( f_{x_1}(\mathbf{x}) \cdots f_{x_n}(\mathbf{x}) \right)^T = 0
\]

and so \( \nabla f(\mathbf{x}) = 0 \). This proves the following theorem.
Theorem 18.1.3 Suppose $U$ is an open set contained in $D(f)$ such that $f$ is $C^1$ on $U$ and suppose $x \in U$ is a local minimum or local maximum for $f$. Then $\nabla f(x) = 0$.

Definition 18.1.4 A singular point for $f$ is a point $x$ where $\nabla f(x) = 0$. This is also called a critical point. By analogy with the one variable case, a point where the gradient does not exist will also be called a critical point.

Example 18.1.5 Find the critical points for the function, $f(x, y) = xy - x - y$ 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?

$$f_x = y - 1 = 0, \quad f_y = x - 1 = 0$$

and so there is exactly one critical point, $(1, 1)$.

Example 18.1.6 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, letting $(x, y)$ take the place of $(x_0, y_0)$ for simplicity, the function to minimize is

$$f(x, y) = \frac{1}{6} xy \sqrt{4 - x^2 - y^2} \frac{64}{(4 - x^2 - y^2)}$$

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

$$\frac{-4 + 2x^2 + y^2}{x^2y(4 - x^2 - y^2)} = 0,$$

and

$$\frac{-4 + x^2 + 2y^2}{xy^2(4 - x^2 - y^2)} = 0.$$  

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. How do you know this is not the largest tetrahedron?

Example 18.1.7 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 \( xyz = 32 \). How do you know this gives the least surface area? Why doesn’t this give the largest surface area?

### 18.2 The Second Derivative Test

#### 18.2.1 Functions Of Two Variables

In the special case of a function of two variables, \( f(x, y) \) which is the only case considered in most calculus books, the second derivative test is given in the following theorem. It is a black box formulation of the second derivative test.

**Theorem 18.2.1 (Second Derivative Test)** Let \( f \) be a function of two variables defined on an open set, \( U \) whose second order partial derivatives exist and are continuous. That is, \( f \in C^2(U) \). Suppose \( (a, b) \in U \) is a point where both partial derivatives of \( f \) vanishes. That is \( f_x(a, b) = f_y(a, b) = 0 \). Let

\[
D = f_{xx}(a, b)f_{yy}(a, b) - (f_{xy}(a, b))^2.
\]

Then

1. If \( D > 0 \) and \( f_{xx}(a, b) < 0 \), then \( f \) has a local maximum at \( (a, b) \).
2. If \( D > 0 \) and \( f_{xx}(a, b) > 0 \), then \( f \) has a local minimum at \( (a, b) \).
3. If \( D < 0 \), then \( f \) has a saddle point at \( (a, b) \).
4. If \( D = 0 \), the test fails.

The above is really a statement about the eigenvalues of the Hessian matrix,

\[
H = \begin{pmatrix}
  f_{xx}(a, b) & f_{x,y}(a, b) \\
  f_{x,y}(a, b) & f_{yy}(a, b)
\end{pmatrix}
\]

at a point \( (a, b) \) where the partial derivatives of \( f \) vanish. It reduces to the following much simpler statement. If both eigenvalues of \( H \) are positive, then \( f \) has a local minimum at \( (a, b) \). If both eigenvalues are negative, then \( f \) has a local maximum at \( (a, b) \). If one eigenvalue is positive and one is negative, then you have a saddle point at \( (a, b) \). If at least one eigenvalue equals zero, then the test fails. Here is a picture which may help you remember this second version of this test.
Use whichever version of this theorem you find easiest to remember. However, in the case of a function of many variables, the description I just gave has an obvious generalization. This is presented next. If you are not interested in it, I think you can skip it because it isn’t included in the book for the course.

### 18.2.2 Functions Of Many Variables

There is a version of the second derivative test for a function of many variables 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 19.4.

**Definition 18.2.2** The matrix, $H(x)$ whose $ij$th entry at the point $x$ is $\frac{\partial^2 f}{\partial x_i \partial x_j}(x)$ is called the Hessian matrix.

The following theorem says that if all the eigenvalues of the Hessian matrix at a critical point are positive, then the critical point is a local minimum. If all the eigenvalues of the Hessian matrix at a critical point are negative, then the critical point is a local maximum. Finally, if some of the eigenvalues of the Hessian matrix at the critical point are positive and some are negative then the critical point is a saddle point. The following picture illustrates the situation.

![Emojis representing the second derivative test results]

**Theorem 18.2.3** Let $f : U \rightarrow \mathbb{R}$ 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, $H(x)$. 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.

Here is an example.

**Example 18.2.4** Let $f(x, y) = 10xy + y^2$. Find the critical points and determine whether they are local minima, local maxima or saddle points.
First
\[ \nabla (10xy + y^2) = (10y, 10x + 2y) \]
and so there is one critical point at the point \((0, 0)\). What is it? The Hessian matrix is
\[
\begin{pmatrix}
0 & 10 \\
10 & 2
\end{pmatrix}
\]
and the eigenvalues are of different signs. Therefore, the critical point \((0, 0)\) is a saddle point. Here is a graph drawn by Maple.

Here is another example.

**Example 18.2.5** 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.

\[ f_x(x, y) = 8x^3 - 12x^2 + 28x + 24yx - 12y - 12 \quad \text{and} \quad f_y(x, y) = 12x^2 - 12x + 4y + 4. \]

The points at which both \( f_x \) and \( f_y \) equal zero are \( \left( \frac{1}{2}, -\frac{1}{4} \right) \), \((0, -1)\), 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 \( \left( \frac{1}{2}, -\frac{1}{4} \right) \). 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. To determine this, find the eigenvalues.

\[
\det \left( \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 4 & -12 \\ -12 & 4 \end{pmatrix} \right) = \lambda^2 - 8\lambda - 128 = (\lambda + 8)(\lambda - 16)
\]
so the eigenvalues are −8 and 16 as claimed.

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.

Below is a graph of this function which illustrates the behavior near saddle points.
18.2. THE SECOND DERIVATIVE TEST

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.

Example 18.2.6 Suppose \( f(x, y) = 6xy^2 - 2x^3 - 3y^4 \). Show that \((0, 0)\) 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

\[
f(x, y) = 6xy^2 - 2x^3 - 3y^4\]

and so \( f_x(0, 0) = f_y(0, 0) = 0 \).

\[
f_x(x, y) = 6y^2 - 6x^2, \quad f_y(x, y) = 12xy - 12y^3\]

and clearly \((0, 0)\) is a critical point. So are \((1, 1)\) and \((1, -1)\). Now \( f_{xx}(0, 0) = 0 \) and so are \( f_{xy}(0, 0) \) and \( f_{yy}(0, 0) \). 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 \( h(t) = f(t, t) = 4t^3 - 3t^4 \), 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 \((0, 0)\).

Example 18.2.7 Find the critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.
\[ f(x, y, z) = \frac{5}{6} x^2 + 4x + 16 - \frac{7}{3} xy - 4y - \frac{4}{3} xz + 12z + \frac{5}{6} y^2 - \frac{4}{3} zy + \frac{1}{3} z^2 \]

First you need to locate the critical points. This involves taking the gradient.

\[
\nabla \left( \frac{5}{6} x^2 + 4x + 16 - \frac{7}{3} xy - 4y - \frac{4}{3} xz + 12z + \frac{5}{6} y^2 - \frac{4}{3} zy + \frac{1}{3} z^2 \right) = \left( \frac{5}{3} x + 4 - \frac{7}{3} y - \frac{4}{3} z, -\frac{7}{3} x - 4 + \frac{5}{3} y - \frac{4}{3} z, -\frac{4}{3} x + 12 - \frac{4}{3} y + \frac{2}{3} z \right)
\]

Next you need to set the gradient equal to zero and solve the equations. This yields \( y = 5, x = 3, z = -2 \). Now to use the second derivative test, you assemble the Hessian matrix which is

\[
\begin{pmatrix}
\frac{5}{3} & -\frac{7}{3} & -\frac{4}{3} \\
-\frac{7}{3} & \frac{5}{3} & -\frac{4}{3} \\
-\frac{4}{3} & -\frac{4}{3} & \frac{2}{3}
\end{pmatrix}
\]

Note that in this simple example, the Hessian matrix is constant and so all that is left is to consider the eigenvalues. Writing the characteristic equation and solving yields the eigenvalues are 2, -2, 4. Thus the given point is a saddle point.

### 18.3 Lagrange Multipliers

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 in the following picture.

![Lagrange Multipliers Diagram](image-url)
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 \( \mathbf{x}(t) = (x(t), y(t), z(t)) \) 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
\[
0 = h'(t_0) = \nabla f(x(t_0), y(t_0), z(t_0)) \cdot \mathbf{x}'(t_0)
\]
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.

In solving the equations which result in the case there are fewer than three variables so that \( \nabla f \) and \( \nabla g \) are in either \( \mathbb{R}^2 \) or \( \mathbb{R}^3 \), there is a convenient observation which often makes it easier to get to the answer. The above says
\[
(f_x, f_y, f_z) = \lambda (g_x, g_y, g_z)
\]
or in the case where there are only two variables,
\[
(f_x, f_y, 0) = \lambda (g_x, g_y, 0)
\]
Since the two vectors are multiples of each other, this says
\[
(f_x, f_y, f_z) \times (g_x, g_y, g_z) = 0, \ (f_x, f_y, 0) \times (g_x, g_y, 0) = 0.
\]
I will illustrate how to use this with the examples.

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 18.3.1** _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\(^*\),
\[
(yz, xz, xy) = \lambda (2x, 2y, 2z).
\]
Therefore, each of \( 2x \lambda x^2, 2y \lambda y^2, 2z \lambda z^2 \) 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.

\(^*\)There exists such a point because the sphere is closed and bounded.
Also, you can observe that

\[ 0 = (yz, xz, xy) \times (2x, 2y, 2z) = (2xz^2 - 2xy^2, 2x^2y - 2yz^2, 2y^2z - 2x^2z) \]

and so, \( z^2 = y^2 \) and \( x^2 = z^2 \) so from the equation of constraint, \( 3x^2 = 27 \) and \( x = \pm 3 \). Now it follows you get the same candidates as before. Note I divided by 2 to get \( z^2 = y^2 \) but this is all right because I am sure that the correct answer does not include \( x = 0 \) or any of the other variables for that matter.

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 18.3.2** 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 \( \left( \frac{1}{2}, -\frac{\sqrt{3}}{2} \right) \).

Using the trick with the cross product,

\[ 0 = (y, x + 1, 0) \times (2x, 2y, 0) = (0, 0, 2y^2 - 2(x + 1)x) \]

and so you see right away that \( y^2 = x^2 + x \). Now plug this in to the constraint equation. to obtain

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

which has solutions \( x = -1, x = \frac{1}{2} \) and so this leads to candidates \((−1, 0), \left( \frac{1}{2}, \frac{\sqrt{3}}{2} \right), \left( \frac{1}{2}, -\frac{\sqrt{3}}{2} \right) \) as before.

**Example 18.3.3** Find candidates for 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
18.3. LAGRANGE MULTIPLIERS

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.

\[
\begin{align*}
  y - 2x - \lambda (2x + 2y) &= 0 \\
  x - 2\lambda (x + y) &= 0 \\
  x^2 + 2xy + y^2 &= 4
\end{align*}
\]

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,

\[
y = 3x \quad (18.1)
\]

From the constraint equation,

\[
x^2 + 2x (3x) + (3x)^2 = 4
\]

and so

\[
x = \frac{1}{2} \text{ or } -\frac{1}{2}
\]

Now from (18.2), the points of interest on the boundary of this set are

\[
\left( \frac{1}{2}, \frac{3}{2} \right), \text{ and } \left( -\frac{1}{2}, -\frac{3}{2} \right).
\]

\[
\begin{align*}
  f \left( \frac{1}{2}, \frac{3}{2} \right) &= \left( \frac{1}{2} \right) \left( \frac{3}{2} \right) - \left( \frac{1}{2} \right)^2 \\
  &= \frac{1}{2}
\end{align*}
\]

\[
\begin{align*}
  f \left( -\frac{1}{2}, -\frac{3}{2} \right) &= \left( -\frac{1}{2} \right) \left( -\frac{3}{2} \right) - \left( -\frac{1}{2} \right)^2 \\
  &= \frac{1}{2}
\end{align*}
\]

It follows the candidates for maximum and minimum are \( \left( \frac{1}{2}, \frac{3}{2} \right), (0,0), \text{ and } \left( -\frac{1}{2}, -\frac{3}{2} \right) \). Therefore it appears that one of these yields a maximum and another the minimum. However, this is a little misleading. How do you even know a maximum or a minimum exists? The set, \( x^2 + 2xy + y^2 \leq 4 \) is an unbounded set which lies between the two lines \( x + y = 2 \) and \( x + y = -2 \). In fact there is no minimum. For example, take \( x = 100, y = -98 \). Then \( xy - x^2 = x (y - x) = 100 (-98 - 100) \) which is a large negative number much less than 0, the answer for the point \((0,0)\).
Now let's use the cross product trick.

\[
\nabla (xy - x^2) = (y - 2x, x), \\
\nabla (x^2 + 2xy + y^2 - 4) = (2x + 2y, 2x + 2y)
\]

Therefore, you need

\[
(y - 2x, x, 0) \times (2x + 2y, 2x + 2y, 0) = (0, 0, -4xy + 2y^2 - 6x^2) = 0
\]

Thus,

\[
y^2 - 2xy - 3x^2 = 0 \quad (18.3)
\]

and you also need

\[
x^2 + 2xy + y^2 = 4. \quad (18.4)
\]

The candidates are the solutions of the two equations. Of course there is no way to solve a system of nonlinear equations which will always work. In this case you can notice that \(18.3\) implies \(x + y = 2\) or \(x + y = -2\). Then if \(x + y = 2\) you could use this in \(18.3\) and write

\[
(x - 2)^2 - 2x (2 - x) - 3x^2 = 0
\]

Then you solve this which yields \(x = \frac{1}{2}\) and so \(y = \frac{3}{2}\). A candidate is \(\left(\frac{1}{2}, \frac{3}{2}\right)\). Now you suppose \(x + y = -2\) and substitute this in to \(18.3\)

\[
(-x - 2)^2 - 2x (-x - 2) - 3x^2 = 0
\]

The only solution is \(x = -\frac{1}{2}\) and this leads to the candidate \(\left(-\frac{1}{2}, -\frac{3}{2}\right)\). These are the same candidates obtained earlier.

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.

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. If you are interested, there is a complete proof in an appendix to this book.

First here is a simple definition explaining one of the terms in the statement of this theorem.

**Definition 18.3.4** 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 18.3.5** Let \(U\) be an open subset of \(\mathbb{R}^n\) and let \(f: U \rightarrow \mathbb{R}\) 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

\[
g_i(x) = 0, \quad i = 1, \cdots, m \quad (18.5)
\]

and if some \(m \times m\) submatrix of

\[
Dg(x_0) = \begin{pmatrix}
ge_{1x_1}(x_0) & g_{1x_2}(x_0) & \cdots & g_{1x_n}(x_0) \\
\vdots & \vdots & \ddots & \vdots \\
g_{mx_1}(x_0) & g_{mx_2}(x_0) & \cdots & g_{mx_n}(x_0)
\end{pmatrix}
\]
has nonzero determinant, then there exist scalars, \( \lambda_1, \cdots, \lambda_m \) such that
\[
\begin{pmatrix}
  f_{x_1}(x_0) \\
  \vdots \\
  f_{x_n}(x_0)
\end{pmatrix}
= \lambda_1
\begin{pmatrix}
  g_{1x_1}(x_0) \\
  \vdots \\
  g_{1x_n}(x_0)
\end{pmatrix}
+ \cdots + \lambda_m
\begin{pmatrix}
  g_{mx_1}(x_0) \\
  \vdots \\
  g_{mx_n}(x_0)
\end{pmatrix}
\tag{18.6}
\]
holds.

To help remember how to use it, 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 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, \( x_1, \cdots, x_n, \lambda_1, \cdots, \lambda_m \). 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 18.3.6** 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 following 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 \( x = 2\sqrt{\frac{3}{5}}, y = \sqrt{\frac{3}{5}}, \) 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 \( (2\sqrt{8/15}, \sqrt{8/15}, \pm \sqrt{4/3}) \), and \( (-2\sqrt{8/15}, -\sqrt{8/15}, \pm \sqrt{4/3}) \), a choice of 4 points to check. Clearly the one which gives the smallest value is

\[
\left( 2\sqrt{8/15}, \sqrt{8/15}, -\sqrt{4/3} \right)
\]

or \( (-2\sqrt{8/15}, -\sqrt{8/15}, -\sqrt{4/3}) \) 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 \( z \).

### 18.4 Finding Function From Its Gradient

Sometimes you have a vector valued function \( g \), referred to as a vector field, and you want to find a scalar valued function \( f \), called a scalar potential, such that \( \nabla f = g \). Here we will mainly consider the case of functions of two variables. Here is an example:

**Example 18.4.1** Here is a vector valued function of two variables

\[
\mathbf{g}(x, y) = \begin{pmatrix} 2xy + 1 \\ x^2 \end{pmatrix}
\]

Find if possible a scalar valued function \( f(x, y) \) such that \( \nabla f(x, y) = \mathbf{g}(x, y) \).

You need

\[
f_x = 2xy + 1
\]

This requires that if such an \( f \) exists, that

\[
f(x, y) = x^2y + x + h(y)
\]

Note how an unknown function of \( y \) was added in rather than a constant. This is because you want to get the most general possible description of \( f \). Otherwise, you might be throwing out the solution. Now differentiate both sides with respect to \( y \) to obtain that

\[
f_y = x^2 + h'(y) = x^2
\]

Hence, in this case \( h'(y) = 0 \) and so in fact the solution is

\[
f(x, y) = x^2y + x
\]
Next check your work to see if this really is a scalar potential by taking its gradient and seeing if it is what it is supposed to be.

Is there a way to tell if a vector valued function of two variables has a scalar potential? In fact there is. One way is to just follow the above procedure. Another way is to use the following condition which is discussed completely later in the book.

**Condition 18.4.2** Let \( g(x, y) = \begin{pmatrix} M(x, y) \\ N(x, y) \end{pmatrix} \), where \( M, N \) are \( C^1 \) functions defined on \( \mathbb{R}^2 \). Then \( g \) has a scalar potential, also referred to as “\( g \) is conservative” if and only if \( N_x = M_y \).

In the above example, you have \( M = 2xy + 1 \) and \( N = x^2 \). Then \( N_x = 2x \) and \( M_y = 2x \) and so a scalar potential will exist and the above process found it. It is easy to see that the above is necessary. If you have

\[
f_x = M, \quad f_y = N
\]

then you have

\[
f_y = M_y, \quad f_{yx} = N_x
\]

and by equality of mixed partial derivatives, you have

\[
M_y = N_x
\]

Conversely, if the condition holds, why does there exist a scalar potential? Try

\[
f(x, y) = \int_0^y N(x, t) \, dt + \int_0^x M(t, 0) \, dt
\]

Does it work? Go ahead and differentiate under the integral sign.

\[
f_x = \int_0^y N_x(x, t) \, dt + M(x, 0)
\]

\[
= \int_0^y M_y(x, t) \, dt + M(x, 0)
\]

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

\[
f_y(x, y) = N(x, y)
\]

so it appears that this is a valid potential function. However, the differentiation under the integral sign has not been justified. However, it will be valid if the functions \( N, M \) are \( C^1 \).

One way this is sometimes written is as follows.

\[
df = f_x dx + f_y dy = Mdx + Ndy
\]

**Example 18.4.3** Find a scalar potential for the vector field

\[
\begin{pmatrix} xy \\ y \end{pmatrix}
\]

In this case, \( M_y = x \) and \( N_x = 0 \) and these are not equal, so there is no scalar potential. If you look real hard, you won’t find it because it is **not there to be found**. However, the above process will lead you to the same conclusion. You want

\[
f_x = xy
\]
and so

\[ f(x, y) = \frac{x^2}{2} y + h(y) \]

Then if the scalar potential \( f \) exists, you would need to have

\[ y = f_y = \frac{x^2}{2} y + h'(y) \]

Hence \( y - h'(y) = \frac{x^2}{2} y \). If you change \( x \), the right side changes but the left side does not. Hence there is no solution to this problem. We say that this vector field is non conservative. We say that the “differential form \( xydx + ydy \) is not exact.

### 18.5 Exercises With Answers

1. Maximize \( x + 3y - 6z \) subject to the constraint, \( x^2 + 2y^2 + z^2 = 9 \).

The Lagrangian is \( L = x + 3y - 6z - \lambda (x^2 + 2y^2 + z^2 - 9) \). Now take the derivative with respect to \( x \). This gives the equation \( 1 - 2\lambda x = 0 \). Next take the derivative with respect to \( y \). This gives the equation \( 3 - 4\lambda y = 0 \). The derivative with respect to \( z \) gives \(-6 - 2\lambda z = 0 \). Clearly \( \lambda \neq 0 \) since this would contradict the first of these equations. Similarly, none of the variables, \( x, y, z \) can equal zero. Solving each of these equations for \( \lambda \) gives \( \lambda = \frac{1}{2} \) and \( \lambda = -\frac{3}{2} \). Thus \( y = \frac{3x}{2} \) and \( z = -6x \). Now you use the constraint equation plugging in these values for \( y \) and \( z \).

\[ x^2 + 2 \left( \frac{3x}{2} \right)^2 + (-6x)^2 = 9 \]

This gives the values for \( x \) as \( x = \frac{3\sqrt{166}}{83}, x = -\frac{3\sqrt{166}}{83} \). From the three equations above, this also determines the values of \( z \) and \( y \).

\[ y = \frac{9\sqrt{166}}{106}, \quad z = -\frac{18\sqrt{166}}{83} \]

Thus there are two points to look at. One will give the minimum value and the other will give the maximum value. You know the minimum and maximum exist because of the extreme value theorem. The two points are \( \left( \frac{3\sqrt{166}}{83}, \frac{9\sqrt{166}}{106}, -\frac{18\sqrt{166}}{83} \right) \) and \( \left( -\frac{3\sqrt{166}}{83}, -\frac{9\sqrt{166}}{106}, \frac{18\sqrt{166}}{83} \right) \). Now you just need to find which is the minimum and which is the maximum. Plug these in to the function you are trying to maximize. \( \left( \frac{3\sqrt{166}}{83} \right)^2 + 3 \left( \frac{9\sqrt{166}}{106} \right) - 6 \left( -\frac{18\sqrt{166}}{83} \right) \) will clearly be the maximum value occurring at \( \left( \frac{3\sqrt{166}}{83}, \frac{9\sqrt{166}}{106}, -\frac{18\sqrt{166}}{83} \right) \). The other point will obviously yield the minimum because this one is positive and the other one is negative. If you use a calculator to compute this you get \( \left( \frac{3\sqrt{166}}{83} \right)^2 + 3 \left( \frac{9\sqrt{166}}{106} \right) - 6 \left( -\frac{18\sqrt{166}}{83} \right) = 19.326 \).

2. Find the dimensions of the largest rectangle which can be inscribed in a the ellipse \( x^2 + 4y^2 = 4 \).

This is one which you could do without Lagrange multipliers. However, it is easier with Lagrange multipliers. Let a corner of the rectangle be at \((x, y)\). Then the area of the rectangle will be \(4xy\) and since \((x, y)\) is on the ellipse, you have the constraint \(x^2 + 4y^2 = 4\). Thus the problem is to maximize \(4xy\) subject to \(x^2 + 4y^2 = 4\). The Lagrangian is then \( L = 4xy - \lambda (x^2 + 4y^2 - 4)\) and so you get the equations \(4y - 2\lambda x = 0\) and \(4x - 8\lambda y = 0\). You can’t have both \(x\) and \(y\) equal to zero and satisfy the constraint. Therefore, the determinant of the matrix of coefficients must equal zero. Thus \(\begin{vmatrix} -2\lambda & 4 \\ 4 & -8\lambda \end{vmatrix} = -16\lambda^2 - 16 = 0\). This is because the system of equations is of the form

\[
\begin{pmatrix}
-2\lambda & 4 \\
4 & -8\lambda
\end{pmatrix}
\begin{pmatrix}
x \\
y
\end{pmatrix}
= 
\begin{pmatrix}
0 \\
0
\end{pmatrix}
\]
If the matrix has an inverse, then the only solution would be \( x = y = 0 \) which as noted above can’t happen. Therefore, \( \lambda = \pm 1 \). First suppose \( \lambda = 1 \). Then the first equation says \( 2y = x \). Pluggin this in to the constraint equation, \( x^2 + x^2 = 4 \) and so \( x = \pm \sqrt{2} \). Therefore, \( y = \pm \frac{\sqrt{2}}{2} \). This yields the dimensions of the largest rectangle to be \( 2\sqrt{2} \times \sqrt{2} \). You can check all the other cases and see you get the same thing in the other cases as well.

3. Maximize \( 2x + y \) subject to the condition that \( \frac{x^2}{4} + y^2 \leq 1 \).

The maximum of this function clearly exists because of the extreme value theorem since the condition defines a closed and bounded set in \( \mathbb{R}^2 \). However, this function does not achieve its maximum on the interior of the given ellipse defined by \( \frac{x^2}{4} + y^2 \leq 1 \) because the gradient of the function which is to be maximized is never equal to zero. Therefore, this function must achieve its maximum on the set \( \frac{x^2}{4} + y^2 = 1 \). Thus you want to maximize \( 2x + y \) subject to \( \frac{x^2}{4} + y^2 = 1 \). This is just like Problem 2. You can finish this.

4. Find the points on \( y^2 x = 16 \) which are closest to \( (0,0) \).

You want to minimize \( x^2 + y^2 \) subject to \( y^2 x = 16 \). Of course you really want to minimize \( \sqrt{x^2 + y^2} \) but the ordered pair which minimized \( x^2 + y^2 \) is the same as the ordered pair which minimize \( \sqrt{x^2 + y^2} \) so it is pointless to drag around the square root. The Lagrangian is \( x^2 + y^2 - \lambda (y^2 x - 16) \). Differentiating with respect to \( x \) and \( y \) gives the equations \( 2x - \lambda y^2 = 0 \) and \( 2y - 2\lambda xy = 0 \). Neither \( x \) nor \( y \) can equal zero and solve the constraint. Therefore, the second equation implies \( \lambda x = 1 \). Hence \( \lambda = \frac{1}{x} = \frac{2x}{y^2} \). Therefore, \( 2x^2 = y^2 \) and so \( 2x^3 = 16 \) and so \( x = 2 \). Therefore, \( y = \pm 2\sqrt{2} \).

The points are \( (2,2\sqrt{2}) \) and \( (2,-2\sqrt{2}) \). They both give the same answer. Note how ad hoc these procedures are. I can’t give you a simple strategy for solving these systems of nonlinear equations by algebra because there is none. Sometimes nothing you do will work.

5. Find points on \( xy = 1 \) farthest from \( (0,0) \) if any exist. If none exist, tell why. What does this say about the method of Lagrange multipliers?

If you graph \( xy = 1 \) you see there is no farthest point. However, there is a closest point and the method of Lagrange multipliers will find this closest point. This shows that the answer you get has to be carefully considered to determine whether you have a maximum or a minimum or perhaps neither.

6. A curve is formed from the intersection of the plane, \( 2x + y + z = 3 \) and the cylinder \( x^2 + y^2 = 4 \). Find the point on this curve which is closest to \( (0,0,0) \).

You want to maximize \( x^2 + y^2 + z^2 \) subject to the two constraints \( 2x + y + z = 3 \) and \( x^2 + y^2 = 4 \). This means the Lagrangian will have two multipliers.

\[
L = x^2 + y^2 + z^2 - \lambda (2x + y + z - 3) - \mu (x^2 + y^2 - 4)
\]

Then this yields the equations \( 2x - 2\lambda - 2\mu x = 0 \), \( 2y - \lambda - 2\mu y \), and \( 2z - \lambda = 0 \). The last equation says \( \lambda = 2z \) and so I will replace \( \lambda \) with \( 2z \) where ever it occurs. This yields \( x - 2z - \mu x = 0, 2y - 2z - 2\mu y = 0 \).

This shows \( x (1 - \mu) = 2y (1 - \mu) \). First suppose \( \mu = 1 \). Then from the above equations, \( z = 0 \) and so the two constraints reduce to \( 2y + x = 3 \) and \( x^2 + y^2 = 4 \) and \( 2y + x = 3 \). The solutions are \( (\frac{4}{5} - \frac{\sqrt{7}}{5}, \frac{6}{5} + \frac{\sqrt{7}}{5}, 0), (\frac{4}{5} + \frac{\sqrt{7}}{5}, \frac{6}{5} - \frac{\sqrt{7}}{5}, 0) \).
The other case is that \( \mu \neq 1 \) in which case \( x = 2y \) and the second constraint yields \( y = \pm \frac{3}{\sqrt{5}} \) and \( x = \pm \frac{3}{\sqrt{5}} \). Now from the first constraint, \( z = -2\sqrt{5} + 3 \) in the case where \( y = \frac{2}{\sqrt{5}} \) and \( z = 2\sqrt{5} + 3 \) in the other case. This yields the points \( \left( \frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}, -2\sqrt{5} + 3 \right) \) and \( \left( -\frac{1}{\sqrt{5}}, -\frac{2}{\sqrt{5}}, 2\sqrt{5} + 3 \right) \). This appears to have exhausted all the possibilities and so it is now just a matter of seeing which of these points gives the best answer. An answer exists because of the extreme value theorem. After all, this constraint set is closed and bounded. The first candidate listed above yields for the answer \( (\frac{1}{3} - \frac{3}{5}\sqrt{11})^2 + (\frac{6}{5} + \frac{1}{3}\sqrt{11})^2 = 4 \). The second candidate listed above yields \( (\frac{3}{5} + \frac{2}{5}\sqrt{11})^2 + (\frac{6}{5} - \frac{2}{3}\sqrt{11})^2 = 4 \) also. Thus these two give equally good results. Now consider the last two candidates. \( (\frac{1}{\sqrt{5}})^2 + (\frac{2}{\sqrt{5}})^2 + (-2\sqrt{5} + 3)^2 = 4 + (-2\sqrt{5} + 3)^2 \) which is larger than 4. Finally the last candidate yields \( (-\frac{4}{\sqrt{5}})^2 + (\frac{2}{\sqrt{5}})^2 + (2\sqrt{5} + 3)^2 = 4 + (2\sqrt{5} + 3)^2 \) also larger than 4. Therefore, there are two points on the curve of intersection which are closest to the origin, \( (\frac{2}{5} - \frac{3}{2}\sqrt{11}, \frac{6}{5} + \frac{1}{3}\sqrt{11}, 0) \) and \( (\frac{3}{5} + \frac{2}{5}\sqrt{11}, \frac{6}{5} - \frac{2}{3}\sqrt{11}, 0). \) Both are a distance of 4 from the origin.

7. Here are two lines. \( \mathbf{x} = (1 + 2t, 2 + t, 3 + t)^T \) and \( \mathbf{x} = (2 + s, 1 + 2s, 1 + 3s)^T \). Find points \( \mathbf{p}_1 \) on the first line and \( \mathbf{p}_2 \) on the second with the property that \( |\mathbf{p}_1 - \mathbf{p}_2| \) 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.

**Hint:** Do you need to use Lagrange multipliers for this?

8. Find the point on \( x^2 + y^2 + z^2 = 1 \) closest to the plane \( x + y + z = 10 \).

You want to minimize \( (x - a)^2 + (y - b)^2 + (z - c)^2 \) subject to the constraints \( a + b + c = 10 \) and \( x^2 + y^2 + z^2 = 1 \). There seem to be a lot of variables in this problem, 6 in all. Start taking derivatives and hope for a miracle. This yields \( 2(x - a) - 2\mu x = 0, 2(y - b) - 2\mu y = 0, 2(z - c) - 2\mu z = 0 \). Also, taking derivatives with respect to \( a, b, \) and \( c \) you obtain \( 2(x - a) + \lambda = 0, 2(y - b) + \lambda = 0, 2(z - c) + \lambda = 0 \). Comparing the first equations in each list, you see \( \lambda = 2\mu x \) and then comparing the second two equations in each list, \( \lambda = 2\mu y \) and similarly, \( \lambda = 2\mu z \). Therefore, if \( \mu \neq 0 \), it must follow that \( x = y = z \). Now you can see by sketching a rough graph that the answer you want has each of \( x, y, \) and \( z \) nonnegative. Therefore, using the constraint for these variables, the point desired is \( \left( \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right) \) which you could probably see was the answer from the sketch. However, this could be made more difficult rather easily such that the sketch won’t help but Lagrange multipliers will.
Chapter 19

The Derivative Of Vector Valued Functions, What Is The Derivative?*

If you are going to do this stuff, you might as well do it right and include the case of vector valued functions. You know everything about matrices at this point for it to all make perfect sense. Therefore, I think it is well worth your time to read this although you are unlikely to see it on a test. It is not any harder than what has been presented. It also tells you what the derivative is. This is essential information if you are going to understand Newton's method for nonlinear systems. It is also essential if you want to read a really good book on continuum mechanics and is needed in many other physical and engineering applications. Also included is a proof of the second derivative test.

Recall the following definition.

Definition 19.0.1 A function, $T$ which maps $\mathbb{R}^n$ to $\mathbb{R}^p$ is called a linear transformation if for every pair of scalars, $a,b$ and vectors, $x,y \in \mathbb{R}^n$, it follows that $T(ax+by) = aT(x) + bT(y)$.

Recall that from the properties of matrix multiplication, it follows that if $A$ is an $n \times p$ matrix, and if $x,y$ are vectors in $\mathbb{R}^n$, then $A(ax+by) = aA(x) + bA(y)$. Thus you can define a linear transformation by multiplying by a matrix. Of course the simplest example is that of a $1 \times 1$ matrix or number. You can think of the number 3 as a linear transformation, $T$ mapping $\mathbb{R}$ to $\mathbb{R}$ according to the rule $Tx = 3x$. It satisfies the properties needed for a linear transformation because $3(ax+by) = a3x + b3y = aTx + bTy$. The case of the derivative of a scalar valued function of one variable is of this sort. You get a number for the derivative. However, you can think of this number as a linear transformation. Of course it is not worth the fuss to do so for a function of one variable but this is the way you must think of it for a function of $n$ variables.

Definition 19.0.2 Let $f : U \to \mathbb{R}^p$ 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 $h = (h_1, \cdots, h_n)^T$, $f(x+h) = f(x) + \sum_{i=1}^{n} v_i h_i + o(h)$.

}\text{(19.1)}
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 \( \mathbb{R}^n \),

\[
Df(x) w = \left( \begin{array}{c|c|c}
\vdots & \vdots & \vdots \\
v_1 & \cdots & v_n \\
\vdots & \vdots & \vdots 
\end{array} \right) w.
\]

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 \( \frac{df}{dx} \) or \( \frac{\partial f}{\partial x} \).

**Theorem 19.0.3** Suppose \( f \) is as given above in 19.1. Then

\[
v_k = \lim_{h \to 0} \frac{f(x+he_k) - f(x)}{h} = \frac{\partial f}{\partial x_k}(x),
\]

the \( k^{th} \) partial derivative.

**Proof:** Let \( h = (0, \cdots, h, 0, \cdots, 0)^T = he_k \) where the \( h \) is in the \( k^{th} \) slot. Then 19.1 reduces to

\[
f(x+h) = f(x) + v_k h + o(h).
\]

Therefore, dividing by \( h \)

\[
\frac{f(x+he_k) - f(x)}{h} = v_k + \frac{o(h)}{h}
\]

and taking the limit,

\[
\lim_{h \to 0} \frac{f(x+he_k) - f(x)}{h} = \lim_{h \to 0} \left( v_k + \frac{o(h)}{h} \right) = v_k
\]

and so, the above limit exists. This proves the theorem.

Let \( f : U \to \mathbb{R}^q \) 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,

\[
\left( \begin{array}{cccc}
\frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_p} \\
\frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_p} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial f_q}{\partial x_1} & \frac{\partial f_q}{\partial x_2} & \cdots & \frac{\partial f_q}{\partial x_p}
\end{array} \right).
\]
This gives the form of the matrix which defines the linear transformation, \( Df(x) \). Thus

\[
f(x + v) = f(x) + J(x)v + o(v)
\]  
(19.2)

and to reiterate, the linear transformation which results by multiplication by this \( q \times p \) matrix is known as the derivative.

Sometimes \( x, y, z \) is written 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 19.0.4 Let \( A \) be a constant \( m \times n \) matrix and consider \( f(x) = Ax. \) Find \( Df(x) \) if it exists.

\[
f(x + h) - f(x) = A(x + h) - A(x) = Ah = Ah + o(h).
\]

In fact in this case, \( o(h) = 0. \) Therefore, \( Df(x) = A. \) Note that this looks the same as the case in one variable, \( f(x) = ax. \)

19.1 \( C^1 \) 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 19.1.1 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 \( Df(0). \)

\[
f(h) - f(0) = 0h + h^2 \sin \left( \frac{1}{h} \right) = o(h)
\]

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 19.1.2 Let \( f(x, y) = \begin{pmatrix} x^2y + y^2 \\ y^3x \end{pmatrix}. \) 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

\[
f(1 + h_1, 2 + h_2) - f(1, 2)
\]
When and this will be the derivative. There is no question of existence for the
Suppose then the function is said to be \( C \) valued functions,
Therefore, the standard matrix of the derivative is \( \left( \begin{array}{c} 4 \\ 8 \\ 5 \\ 12 \end{array} \right) \).
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 19.1.3** 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 \ref{thm:derivative_existence} 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 theorem.
**Theorem 19.1.4** Suppose \( f : U \to \mathbb{R}^p \) where \( U \) is an open set in \( \mathbb{R}^n \). Suppose also
that all partial derivatives of \( f \) exist on \( U \) and are continuous. Then \( f \) is differentiable at
every point of \( U \).
**Proof:** If you fix all the variables but one, you can apply the fundamental theorem of
calculus as follows.
\[
f(x + v_k e_k) - f(x) = \int_0^1 \frac{\partial f}{\partial x_k} (x + tv_k e_k) v_k dt. \tag{19.3}
\]
Here is why. Let \( h(t) = f(x + tv_k e_k) \). Then
\[
\frac{h(t + h) - h(t)}{h} = \frac{f(x + tv_k e_k + hv_k e_k) - f(x + tv_k e_k)}{hv_k} v_k
\]
and so, taking the limit as \( h \to 0 \) yields
\[
h'(t) = \frac{\partial f}{\partial x_k} (x + tv_k e_k) v_k
\]
Therefore,
\[
f(x + v_k e_k) - f(x) = h(1) - h(0) = \int_0^1 h'(t) dt = \int_0^1 \frac{\partial f}{\partial x_k} (x + tv_k e_k) v_k dt.
\]
Now I will use this observation to prove the theorem. Let \( v = (v_1, \ldots, v_n) \) with \( |v| \)
sufficiently small. Thus \( v = \sum_{k=1}^n v_k e_k \). For the purposes of this argument, define
\[
\sum_{k=n+1}^n v_k e_k = 0.
\]
Then with this convention,

\[
\begin{align*}
\| x + v - x \| & = \sum_{i=1}^{n} \left( f \left( x + \sum_{k=i}^{n} v_k e_k \right) - f \left( x + \sum_{k=i+1}^{n} v_k e_k \right) \right) \\
& = \sum_{i=1}^{n} \int_{0}^{1} \frac{\partial f}{\partial x_i} \left( x + \sum_{k=i+1}^{n} v_k e_k + tv_i e_i \right) v_i dt \\
& = \sum_{i=1}^{n} \int_{0}^{1} \left( \frac{\partial f}{\partial x_i} \left( x + \sum_{k=i+1}^{n} v_k e_k + tv_i e_i \right) v_i - \frac{\partial f}{\partial x_i} (x) v_i \right) dt \\
& \quad + \sum_{i=1}^{n} \int_{0}^{1} \frac{\partial f}{\partial x_i} (x) v_i dt \\
& = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} (x) v_i \\
& \quad + \int_{0}^{1} \sum_{i=1}^{n} \left( \frac{\partial f}{\partial x_i} \left( x + \sum_{k=i+1}^{n} v_k e_k + tv_i e_i \right) - \frac{\partial f}{\partial x_i} (x) \right) v_i dt \\
& = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} (x) v_i + o(v)
\end{align*}
\]

and this shows \( f \) is differentiable at \( x \).

Some explanation of the step to the last line is in order. The messy thing at the end is \( o(v) \) because of the continuity of the partial derivatives. In fact, from the Cauchy Schwarz inequality,

\[
\int_{0}^{1} \sum_{i=1}^{n} \left( \frac{\partial f}{\partial x_i} \left( x + \sum_{k=i+1}^{n} v_k e_k + tv_i e_i \right) - \frac{\partial f}{\partial x_i} (x) \right) v_i dt \leq \int_{0}^{1} \left( \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \left( x + \sum_{k=i+1}^{n} v_k e_k + tv_i e_i \right) - \frac{\partial f}{\partial x_i} (x) \right|^2 \right)^{1/2} dt \left( \sum_{i=1}^{n} v_i^2 \right)^{1/2}
\]

\[
= \int_{0}^{1} \left( \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \left( x + \sum_{k=i+1}^{n} v_k e_k + tv_i e_i \right) - \frac{\partial f}{\partial x_i} (x) \right|^2 \right)^{1/2} dt |v|
\]

Thus, dividing by \( |v| \) and taking a limit as \( |v| \to 0 \), the quotient is nothing but

\[
\int_{0}^{1} \left( \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \left( x + \sum_{k=i+1}^{n} v_k e_k + tv_i e_i \right) - \frac{\partial f}{\partial x_i} (x) \right|^2 \right)^{1/2} dt
\]

which converges to 0 due to continuity of the partial derivatives of \( f \). This proves the theorem.

Here is an example to illustrate.

**Example 19.1.5** Let \( f(x, y) = \left( x^2 y + y^2 \right) \). Find \( Df(x, y) \).
CHAPTER 19. THE DERIVATIVE OF VECTOR VALUED FUNCTIONS, WHAT IS THE DERIVATIVE?

From Theorem 19.1.4 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 19.1.6** Let \( f(x_1, x_2, x_3) = \begin{pmatrix} x_1^2 x_2^2 + x_2^2 \\ x_2 x_1 + x_3 \\ \sin(x_1 x_2 x_3) \end{pmatrix} \). 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

\[
\begin{pmatrix}
2x_1 x_2 & x_2^2 + 2x_2 & 0 \\
x_2 & x_1 & 1 \\
x_2 x_3 \cos(x_1 x_2 x_3) & x_1 x_3 \cos(x_1 x_2 x_3) & x_1 x_2 \cos(x_1 x_2 x_3)
\end{pmatrix}
\]

**Example 19.1.7** 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 \((1, 2, 3)\),

\[
f(x, y, z) \approx f(1, 2, 3) + 1(x - 1) + 2(y - 2) + 6(z - 3) = 11 + 1(x - 1) + 2(y - 2) + 6(z - 3) = -12 + x + 2y + 6z
\]

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 19.1.8** 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 \( C \geq \max \left\{ \left| \frac{\partial f}{\partial x_i}(x_0) \right|, i = 1, \cdots, p \right\} \), then whenever \( |x - x_0| \) is small enough,

\[
|f(x) - f(x_0)| \leq (Cp + 1) |x - x_0|
\]

**Proof:** Suppose \( f \) is differentiable. Since \( o(v) \) satisfies \( \|v\| \) there exists \( \delta_1 > 0 \) such that if \( |x - x_0| < \delta_1 \), then \( |o(x - x_0)| < |x - x_0| \). But also,

\[
\left| \sum_{i=1}^p \frac{\partial f}{\partial x_i}(x_0)(x_i - x_{0i}) \right| \leq C \sum_{i=1}^p |x_i - x_{0i}| \leq Cp |x - x_0|
\]
19.2. THE CHAIN RULE

19.2.1 The Chain Rule For Functions Of One Variable

First recall the chain rule for a function of one variable. Consider the following picture.

\[ I \xrightarrow{f} J \xrightarrow{g} \mathbb{R} \]

Here \( I \) and \( J \) are open intervals and it is assumed that \( g(I) \subseteq J \). The chain rule says that if \( f'(g(x)) \) exists and \( g'(x) \) exists for \( x \in I \), then the composition, \( f \circ g \) also has a derivative at \( x \) and

\[ (f \circ g)'(x) = f'(g(x))g'(x). \]

Recall that \( f \circ g \) is the name of the function defined by \( f \circ g(x) \equiv f(g(x)) \). In the notation of this chapter, the chain rule is written as

\[ Df(g(x))Dg(x) = D(f \circ g)(x). \quad (19.5) \]

19.2.2 The Chain Rule For Functions Of Many Variables

Let \( U \subseteq \mathbb{R}^n \) and \( V \subseteq \mathbb{R}^p \) be open sets and let \( f \) be a function defined on \( V \) having values in \( \mathbb{R}^q \) while \( g \) is a function defined on \( U \) such that \( g(U) \subseteq V \) as in the following picture.

\[ U \xrightarrow{f} V \xrightarrow{g} \mathbb{R}^q \]

The chain rule says that if the linear transformations (matrices) on the left in \( 19.5 \) both exist then the same formula holds in this more general case. Thus

\[ Df(g(x))Dg(x) = D(f \circ g)(x) \]

Note this all makes sense because \( Df(g(x)) \) is a \( q \times p \) matrix and \( Dg(x) \) is a \( p \times n \) matrix. Remember it is all right to do \( (q \times p)(p \times n) \). The middle numbers match. More precisely,

**Theorem 19.2.1** (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,

\[ D(f \circ g)(x) = Df(g(x))Dg(x). \quad (19.6) \]

In particular,

\[ \frac{\partial (f \circ g)(x)}{\partial x_j} = \sum_{i=1}^{p} \frac{\partial f(g(x))}{\partial y_i} \frac{\partial y_i}{\partial x_j}. \quad (19.7) \]
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
\[
\frac{\partial w}{\partial y_i} \frac{\partial y_i}{\partial x_k} = \frac{\partial z}{\partial x_k}.
\]

**Example 19.2.2** 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 (19.3),
\[
\frac{\partial z}{\partial t} = \frac{\partial z}{\partial u} \frac{\partial u}{\partial t} + \frac{\partial z}{\partial v} \frac{\partial v}{\partial t} = v \cos(uv) \sin(x) + u \cos(uv).
\]

Here \( y_1 = u, y_2 = v, t = x_k \). Also,
\[
\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u} \frac{\partial u}{\partial x} + \frac{\partial z}{\partial v} \frac{\partial v}{\partial x} = v \cos(uv) t \cos(x) + u \cos^2(x) \cos(uv).
\]

Clearly you can continue in this way taking partial derivatives with respect to any of the other variables.

**Example 19.2.3** Let \( w = f(u_1,u_2) = u_2 \sin(u_1) \) and \( u_1 = x^2y + z, u_2 = \sin(xy) \). Find \( \frac{\partial w}{\partial x} \), \( \frac{\partial w}{\partial y} \), and \( \frac{\partial w}{\partial z} \).

The derivative of \( f \) is of the form \((w_x, w_y, w_z)\) and so it suffices to find the derivative of \( f \) using the chain rule. You need to find \( Df(u_1,u_2) Dg(x,y,z) \) where \( g(x,y) = \left( \begin{array}{c} x^2y + z \\ \sin(xy) \end{array} \right) \). Then \( Dg(x,y,z) = \left( \begin{array}{cc} 2xy & x^2 \\ y \cos(xy) & x \cos(xy) \end{array} \right) \). Also \( Df(u_1,u_2) = (u_2 \cos(u_1), \sin(u_1)) \).

Therefore, the derivative is
\[
Df(u_1,u_2) Dg(x,y,z) = (u_2 \cos(u_1), \sin(u_1)) \left( \begin{array}{cc} 2xy & x^2 \\ y \cos(xy) & x \cos(xy) \end{array} \right) \left( \begin{array}{c} 1 \\ 0 \end{array} \right)
\]

\[
=(2u_2 \cos(u_1) xy + (\sin(u_1)) y \cos xy, u_2 \cos(u_1) x^2 + (\sin(u_1)) x \cos xy, u_2 \cos(u_1)) = (w_x, w_y, w_z)
\]

Thus \( \frac{\partial w}{\partial x} = 2u_2 \cos(u_1) xy + (\sin(u_1)) y \cos xy = 2(\sin(xy)) (x^2 y + z) xy + (\sin(xy)) y \cos xy \) . Similarly, you can find the other partial derivatives of \( w \) in terms of substituting in for \( u_1 \) and \( u_2 \) in the above. Note
\[
\frac{\partial w}{\partial x} = \frac{\partial w}{\partial u_1} \frac{\partial u_1}{\partial x} + \frac{\partial w}{\partial u_2} \frac{\partial u_2}{\partial x}.
\]

In fact, in general if you have \( w = f(u_1,u_2) \) and \( g(x,y,z) = \left( \begin{array}{c} u_1(x,y,z) \\ u_2(x,y,z) \end{array} \right) \), then
\[
D(f \circ g)(x,y,z) \text{ is of the form }
\]
\[
\left( \begin{array}{cc} u_{1x} & u_{1y} \\ u_{2x} & u_{2y} \end{array} \right) \left( \begin{array}{c} w_{u_1} \\ w_{u_2} \end{array} \right)
\]

\[
= \left( \begin{array}{c} w_{u_1} u_x + w_{u_2} u_{2x} + w_{u_1} u_y + w_{u_2} u_{2y} + w_{u_1} u_z + w_{u_2} u_{2z} \end{array} \right).
\]
Example 19.2.4 Let \( w = f(u_1, u_2, u_3) = u_1^2 + u_3 + u_2 \) and \( g(x, y, z) = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} x + 2yz \\ x^2 + y \\ z^2 + x \end{pmatrix} \). Find \( \frac{\partial w}{\partial x} \) and \( \frac{\partial w}{\partial y} \).

By the chain rule,

\[
(w_x, w_y, w_z) = \begin{pmatrix} w_{u_1} & w_{u_2} & w_{u_3} \end{pmatrix} \begin{pmatrix} u_{1x} & u_{1y} & u_{1z} \\ u_{2x} & u_{2y} & u_{2z} \\ u_{3x} & u_{3y} & u_{3z} \end{pmatrix}
\]

\[
= \begin{pmatrix} w_{u_1} u_{1x} + w_{u_2} u_{2x} + w_{u_3} u_{3x} \\ w_{u_1} u_{1y} + w_{u_2} u_{2y} + w_{u_3} u_{3y} \\ w_{u_1} u_{1z} + w_{u_2} u_{2z} + w_{u_3} u_{3z} \end{pmatrix}
\]

Note the pattern.

\[
w_x = w_{u_1} u_{1x} + w_{u_2} u_{2x} + w_{u_3} u_{3x},
\]

\[
w_y = w_{u_1} u_{1y} + w_{u_2} u_{2y} + w_{u_3} u_{3y},
\]

\[
w_z = w_{u_1} u_{1z} + w_{u_2} u_{2z} + w_{u_3} u_{3z}.
\]

Therefore,

\[
w_x = 2u_1 (1) + 1 (2x) + 1 (1) = 2(x + 2yz) + 2x + 1 = 4x + 4yz + 1
\]

and

\[
w_z = 2u_1 (2y) + 1 (0) + 1 (2z) = 4(x + 2yz) y + 2z = 4yx + 8y^2z + 2z.
\]

Of course to find all the partial derivatives at once, you just use the chain rule. Thus you would get

\[
\begin{pmatrix} w_x & w_y & w_z \end{pmatrix} = \begin{pmatrix} 2u_1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2z & 2y \\ 2x & 1 & 0 \\ 1 & 0 & 2z \end{pmatrix}
\]

\[
= \begin{pmatrix} 2u_1 + 2x + 1 & 4u_1z + 1 & 4u_1y + 2z \\ 4x + 4yz + 1 & 4zx + 8yz^2 + 1 & 4yx + 8y^2z + 2z \end{pmatrix}
\]

Example 19.2.5 Let \( f(u_1, u_2) = \begin{pmatrix} u_1^2 + u_2 \\ \sin(u_2) + u_1 \end{pmatrix} \) and \( g(x_1, x_2, x_3) = \begin{pmatrix} u_1(x_1, x_2, x_3) \\ u_2(x_1, x_2, x_3) \end{pmatrix} = \begin{pmatrix} x_1x_2 + x_3 \\ x_2^2 + x_1 \end{pmatrix} \). Find \( D(f \circ g)(x_1, x_2, x_3) \).

To do this,

\[
Df(u_1, u_2) = \begin{pmatrix} 2u_1 & 1 \\ 1 & \cos(u_2) \end{pmatrix},
Dg(x_1, x_2, x_3) = \begin{pmatrix} x_2 & x_1 & 1 \\ x_1 & x_2 & 0 \end{pmatrix}.
\]

Then

\[
Df(g(x_1, x_2, x_3)) = \begin{pmatrix} 2(x_1x_2 + x_3) & 1 \\ 1 & \cos(x_2^2 + x_1) \end{pmatrix}
\]
and so by the chain rule,

\[
D (f \circ g) (x_1, x_2, x_3) = \begin{pmatrix}
Df(g(x)) \\
Dg(x)
\end{pmatrix}
\begin{pmatrix}
2 (x_1 x_2 + x_3) & 1 & 1 \\
1 & \cos (x_2^2 + x_1) & 1 \\
x_2 & x_1 & 1 \\
x_2 + \cos (x_2^2 + x_1) & x_1 + 2x_2 & \cos (x_2^2 + x_1)
\end{pmatrix}
\begin{pmatrix}
x_2 \\
x_1 \\
1 \\
2x_2
\end{pmatrix}
\begin{pmatrix}
1 \\
2x_2 \\
0
\end{pmatrix}
\]

Therefore, in particular,

\[
\frac{\partial f_1 \circ g}{\partial x_1} (x_1, x_2, x_3) = (2x_1 x_2 + 2x_3) x_2 + 1,
\]

\[
\frac{\partial f_2 \circ g}{\partial x_3} (x_1, x_2, x_3) = 1, \quad \frac{\partial f_2 \circ g}{\partial x_2} (x_1, x_2, x_3) = x_1 + 2x_2 (\cos (x_2^2 + x_1)),
\]

e tc.

In different notation, let \( z_1 \\ z_2 = f(u_1, u_2) = \begin{pmatrix} u^2 + u_2 \\ \sin (u_2) + u_1 \end{pmatrix} \). Then

\[
\frac{\partial z_1}{\partial x_1} = \frac{\partial z_2}{\partial x_1} = 2u_1 + 1 = 2(x_1 x_2 + x_3) x_2 + 1.
\]

Example 19.2.6 Let \( f(u_1, u_2, u_3) = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} u^2 + u_2 u_3 \\ u_1^2 + u_2^3 \\ \ln (1 + u_3^2) \end{pmatrix} \) and let \( g(x_1, x_2, x_3, x_4) = \begin{pmatrix} x_1 + x_2^2 + \sin (x_3) + \cos (x_4) \\ x_4^2 - x_1 \\ x_3^2 + x_4 \end{pmatrix} \). Find \( (f \circ g)'(x) \).

\[
Df(u) = \begin{pmatrix}
2u_1 & u_3 & u_2 \\
2u_1 & 3u_2^2 & 0 \\
0 & 0 & \frac{2u_3}{(1 + u_3^2)}
\end{pmatrix}
\]

Similarly,

\[
Dg(x) = \begin{pmatrix}
1 & 2x_2 & \cos (x_3) & -\sin (x_4) \\
-1 & 0 & 0 & 2x_4 \\
0 & 0 & 2x_3 & 1
\end{pmatrix}
\]

Then by the chain rule, \( D (f \circ g) (x) = Df (u) Dg (x) \) where \( u = g (x) \) as described above. Thus \( D (f \circ g) (x) = \)

\[
\begin{pmatrix}
2u_1 & u_3 & u_2 \\
2u_1 & 3u_2^2 & 0 \\
0 & 0 & \frac{2u_3}{(1 + u_3^2)}
\end{pmatrix}
\begin{pmatrix}
1 & 2x_2 & \cos (x_3) & -\sin (x_4) \\
-1 & 0 & 0 & 2x_4 \\
0 & 0 & 2x_3 & 1
\end{pmatrix}
\]

\[
= \begin{pmatrix}
2u_1 - u_3 & 4u_1 x_2 & 2u_1 \cos x_3 + 2u_2 x_3 & -2u_1 \sin x_4 + 2u_3 x_4 + u_2 \\
2u_1 - 3u_2^2 & 4u_1 x_2 & 2u_1 \cos x_3 & -2u_1 \sin x_4 + 6u_2^2 x_4 \\
0 & 0 & 4u_1 \frac{u_3}{1 + u_3^2} x_3 & 2u_1 \frac{u_3}{1 + u_3^2}
\end{pmatrix}
\]

(19.9)
where each \( u_i \) is given by the above formulas. Thus \( \frac{\partial z_1}{\partial x_1} \) equals

\[
2u_1 - u_3 = 2(x_1 + x_2^2 + \sin(x_3) + \cos(x_4)) - (x_3^2 + x_4)
\]

\[
= 2x_1 + 2x_2^2 + 2\sin x_3 + 2\cos x_4 - x_3^2 - x_4.
\]

while \( \frac{\partial z_2}{\partial x_4} \) equals

\[
-2u_1 \sin x_4 + 6u_2 x_4 = -2(x_1 + x_2^2 + \sin(x_3) + \cos(x_4)) \sin(x_4) + 6(x_4^2 - x_1)^2 x_4.
\]

If you wanted \( \frac{\partial z}{\partial x_2} \) it would be the second column of the above matrix in \( \text{19.9} \). Thus \( \frac{\partial z}{\partial x_2} \) equals

\[
\begin{pmatrix}
\frac{\partial z_1}{\partial x_2} \\
\frac{\partial z_2}{\partial x_2} \\
\frac{\partial z_3}{\partial x_2}
\end{pmatrix} =
\begin{pmatrix}
4u_1 x_2 \\
4u_1 x_2 \\
0
\end{pmatrix} =
\begin{pmatrix}
4(x_1 + x_2^2 + \sin(x_3) + \cos(x_4)) x_2 \\
4(x_1 + x_2^2 + \sin(x_3) + \cos(x_4)) x_2 \\
0
\end{pmatrix}.
\]

I hope that by now it is clear that all the information you could desire about various partial derivatives is available and it all reduces to matrix multiplication and the consideration of entries of the matrix obtained by multiplying the two derivatives.

### 19.2.3 The Derivative Of The Inverse Function

**Example 19.2.7** 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 \( \text{19.4.1} \) on Page \( 381 \) \( DI = I \). By the chain rule,

\[
I = DI = Df^{-1}(f(x)) (Df(x)).
\]

Therefore,

\[
Df(x)^{-1} = Df^{-1}(f(x)).
\]

This is equivalent to

\[
Df(f^{-1}(y))^{-1} = Df^{-1}(y)
\]

or

\[
Df(x)^{-1} = Df^{-1}(y), y = f(x).
\]

This is just like a similar situation for functions of one variable. Remember \( (f^{-1})' (f(x)) = 1/f'(x) \). 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

\[
\delta_{ij} = \frac{\partial x_i}{\partial y_k} (f(x)) \frac{\partial y_k}{\partial x_j} (x).
\]

### 19.2.4 Acceleration In Spherical Coordinates

This is an interesting example which can be done with more elegance in a more general setting. However, the more general approach also depends on the chain rule and this is what it is all about, giving examples of the use of the chain rule. Read it if it interests you.
Example 19.2.8 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 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
\[ r(t) = \rho(t)e_\rho(t) = (x(t), y(t), z(t))^T \]
Now this implies the velocity is
\[ r'(t) = \rho'(t)e_\rho(t) + \rho(t)(e_\rho(t))'. \quad (19.10) \]
You see, \( e_\rho = e_\rho(\rho, \theta, \phi) \) where each of these variables is a function of \( t \).
\[ \frac{\partial e_\rho}{\partial \phi} = (\cos \phi \cos \theta, \cos \phi \sin \theta, -\sin \phi)^T = \frac{1}{\rho}e_\phi, \]
\[ \frac{\partial e_\rho}{\partial \theta} = (-\sin \phi \sin \theta, \sin \phi \cos \theta, 0)^T = \frac{1}{\rho}e_\theta, \]
and
\[ \frac{\partial e_\rho}{\partial \rho} = 0. \]
Therefore, by the chain rule,
\[ \frac{de_\rho}{dt} = \frac{\partial e_\rho}{\partial \phi} \frac{d\phi}{dt} + \frac{\partial e_\rho}{\partial \theta} \frac{d\theta}{dt} + \frac{\partial e_\rho}{\partial \rho} \frac{d\rho}{dt} \]
By (19.10),
\[ r' = \rho' e_\rho + \frac{d\phi}{dt} e_\phi + \frac{d\theta}{dt} e_\theta. \quad (19.11) \]
Now things get interesting. This must be differentiated with respect to \( t \). To do so,
\[
\frac{\partial \mathbf{e}_\theta}{\partial \theta} = (-\rho \sin \phi \cos \theta, -\rho \sin \phi \sin \theta, 0)^T = ?
\]
where it is desired to find \( a, b, c \) such that \( ? = a \mathbf{e}_\theta + b \mathbf{e}_\phi + c \mathbf{e}_\rho \). Thus
\[
\begin{pmatrix}
-\rho \sin \phi \sin \theta & \rho \cos \phi \cos \theta & \sin \phi \cos \theta \\
\rho \sin \phi \cos \theta & \rho \cos \phi \sin \theta & \sin \phi \sin \theta \\
0 & -\rho \sin \phi & \cos \phi
\end{pmatrix}
\begin{pmatrix}
a \\
b \\
c
\end{pmatrix}
= \begin{pmatrix}
-\rho \sin \phi \cos \theta \\
-\rho \sin \phi \sin \theta \\
0
\end{pmatrix}
\]
Using Cramer’s rule, the solution is \( a = 0, b = -\cos \phi \sin \phi, \) and \( c = -\rho \sin^2 \phi \). Thus
\[
\frac{\partial \mathbf{e}_\theta}{\partial \theta} = (-\rho \sin \phi \cos \theta, -\rho \sin \phi \sin \theta, 0)^T = (-\cos \phi \sin \phi) \mathbf{e}_\phi + (-\rho \sin^2 \phi) \mathbf{e}_\rho.
\]
Also,
\[
\frac{\partial \mathbf{e}_\theta}{\partial \phi} = (-\rho \cos \phi \sin \theta, \rho \cos \phi \cos \theta, 0)^T = (\cot \phi) \mathbf{e}_\theta
\]
and
\[
\frac{\partial \mathbf{e}_\theta}{\partial \rho} = (-\sin \phi \sin \theta, \sin \phi \cos \theta, 0)^T = \frac{1}{\rho} \mathbf{e}_\theta.
\]
Now in [19.11] it is also necessary to consider \( \mathbf{e}_\phi \).
\[
\frac{\partial \mathbf{e}_\phi}{\partial \phi} = (-\rho \sin \phi \cos \theta, -\rho \sin \phi \sin \theta, -\rho \cos \phi)^T = -\rho \mathbf{e}_\rho
\]
\[
\frac{\partial \mathbf{e}_\phi}{\partial \theta} = (-\rho \cos \phi \sin \theta, \rho \cos \phi \cos \theta, 0)^T = (\cot \phi) \mathbf{e}_\theta
\]
and finally,
\[
\frac{\partial \mathbf{e}_\phi}{\partial \rho} = (\cos \phi \cos \theta, \cos \phi \sin \theta, -\sin \phi)^T = \frac{1}{\rho} \mathbf{e}_\phi.
\]
With these formulas for various partial derivatives, the chain rule is used to obtain \( \mathbf{r}'' \) which will yield a formula for the acceleration in terms of the spherical coordinates and these special vectors. By the chain rule,
\[
\frac{d}{dt} (\mathbf{e}_\rho) = \frac{\partial \mathbf{e}_\rho}{\partial \theta} \dot{\theta}' + \frac{\partial \mathbf{e}_\rho}{\partial \phi} \dot{\phi}' + \frac{\partial \mathbf{e}_\rho}{\partial \rho} \dot{\rho}'
\]
\[
= \frac{\dot{\theta}'}{\rho} \mathbf{e}_\theta + \frac{\dot{\phi}'}{\rho} \mathbf{e}_\phi
\]
\[
\frac{d}{dt} (\mathbf{e}_\theta) = \frac{\partial \mathbf{e}_\theta}{\partial \theta} \dot{\theta}' + \frac{\partial \mathbf{e}_\theta}{\partial \phi} \dot{\phi}' + \frac{\partial \mathbf{e}_\theta}{\partial \rho} \dot{\rho}'
\]
\[
= \dot{\theta}' ((-\cos \phi \sin \phi) \mathbf{e}_\phi + (-\rho \sin^2 \phi) \mathbf{e}_\rho) + \phi' (\cot \phi \mathbf{e}_\theta) + \frac{\dot{\rho}'}{\rho} \mathbf{e}_\rho
\]
\[
\frac{d}{dt} (\mathbf{e}_\phi) = \frac{\partial \mathbf{e}_\phi}{\partial \theta} \dot{\theta}' + \frac{\partial \mathbf{e}_\phi}{\partial \phi} \dot{\phi}' + \frac{\partial \mathbf{e}_\phi}{\partial \rho} \dot{\rho}'
\]
\[
= (\dot{\theta}' \cot \phi) \mathbf{e}_\theta + \phi' (-\rho \mathbf{e}_\rho) + \left( \frac{\dot{\rho}'}{\rho} \mathbf{e}_\phi \right)
\]
By \[ 19.11 \]

\[
\mathbf{r}'' = \rho'' \mathbf{e}_\rho + \phi'' \mathbf{e}_\phi + \theta'' \mathbf{e}_\theta + \rho' \left( \frac{\theta'}{\rho} \mathbf{e}_\theta + \frac{\phi'}{\rho} \mathbf{e}_\phi \right) + \\
\phi' \left( (\theta' \cot \phi) \mathbf{e}_\theta + \phi' (-\rho \mathbf{e}_\rho) + \left( \frac{\phi'}{\rho} \mathbf{e}_\phi \right) \right) + \\
\theta' \left( (\phi' \sin \phi) \mathbf{e}_\phi + (\rho \sin^2 \phi) \mathbf{e}_\rho + \phi' (\cot \phi) \mathbf{e}_\theta + \frac{\rho'}{\rho} \mathbf{e}_\theta \right)
\]

and from the above, this equals

\[
\rho'' \mathbf{e}_\rho + \phi'' \mathbf{e}_\phi + \theta'' \mathbf{e}_\theta + \rho' \left( \frac{\theta'}{\rho} \mathbf{e}_\theta + \frac{\phi'}{\rho} \mathbf{e}_\phi \right) + \\
\phi' \left( (\theta' \cot \phi) \mathbf{e}_\theta + \phi' (-\rho \mathbf{e}_\rho) + \left( \frac{\phi'}{\rho} \mathbf{e}_\phi \right) \right) + \\
\theta' \left( (\phi' \cos \phi + \sin \phi) \mathbf{e}_\phi + (\rho \sin^2 \phi) \mathbf{e}_\rho + \phi' (\cot \phi) \mathbf{e}_\theta + \frac{\rho'}{\rho} \mathbf{e}_\theta \right)
\]

and now all that remains is to collect the terms. Thus \( \mathbf{r}'' \) equals

\[
\mathbf{r}'' = \left( \rho'' - \rho \left( \phi' \right)^2 - \rho \left( \theta' \right)^2 \sin^2 \phi \right) \mathbf{e}_\rho + \left( \phi'' + 2 \frac{\rho' \phi'}{\rho} - \left( \theta' \right)^2 \cos \phi \sin \phi \right) \mathbf{e}_\phi + \\
+ \left( \theta'' + 2 \frac{\theta' \rho'}{\rho} + 2 \phi' \theta' \cot \phi \right) \mathbf{e}_\theta.
\]

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,

\[
\rho'' - \rho \left( \phi' \right)^2 - \rho \left( \theta' \right)^2 \sin^2 \phi = a(t), \\
\phi'' + 2 \frac{\rho' \phi'}{\rho} - \left( \theta' \right)^2 \cos \phi \sin \phi = 0, \\
\theta'' + 2 \frac{\theta' \rho'}{\rho} + 2 \phi' \theta' \cot \phi = 0
\]

along with initial conditions, \( \rho(0) = \rho_0 \) (the distance from the launch site to the center of the earth), \( \rho'(0) = \rho_1 \) (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 \( \mathbf{e}_\theta \) and \( \mathbf{e}_\phi \) 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.

---

\[ ^1 \] You won’t be able to find the solution to equations like these in terms of simple functions. The existence of such functions is being assumed. The reason they exist often depends on the implicit function theorem, a big theorem in advanced calculus.
19.3 Proof Of 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 19.3.1** 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

$$\lim_{v \to 0} \frac{|o(g(x + v) - g(x))|}{|v|} = 0. \tag{19.12}$$

From Lemma 19.1.8, there exists $\delta > 0$ such that if $|v| < \delta$, then

$$|g(x + v) - g(x)| \leq (Cn + 1)|v|. \tag{19.13}$$

Now let $\varepsilon > 0$ be given. There exists $\eta > 0$ such that if $|g(x + v) - g(x)| < \eta$, then

$$|o(g(x + v) - g(x))| < \left(\frac{\varepsilon}{Cn + 1}\right)|g(x + v) - g(x)| \tag{19.14}$$

Let $|v| < \min\left(\delta, \frac{\eta}{Cn + 1}\right)$. For such $v$, $|g(x + v) - g(x)| \leq \eta$, which implies

$$|o(g(x + v) - g(x))| < \left(\frac{\varepsilon}{Cn + 1}\right)|g(x + v) - g(x)| \leq \left(\frac{\varepsilon}{Cn + 1}(Cn + 1)|v|$$

and so

$$\frac{|o(g(x + v) - g(x))|}{|v|} < \varepsilon$$

which establishes (19.12). 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 19.3.2** (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 $f : V \to \mathbb{R}$. Suppose $Dg(x)$ exists for some $x \in U$ and that $Df(g(x))$ exists. Then $D(f \circ g)(x)$ exists and furthermore,

$$D(f \circ g)(x) = Df(g(x))Dg(x). \tag{19.15}$$

In particular,

$$\frac{\partial (f \circ g)}{\partial x_j}(x) = \sum_{i=1}^{p} \frac{\partial f}{\partial y_i}(g(x)) \frac{\partial g_i}{\partial x_j}(x). \tag{19.16}$$

**Proof:** From the assumption that $Df(g(x))$ exists,

$$f(g(x + v)) = f(g(x)) + \sum_{i=1}^{p} \frac{\partial f}{\partial y_i}(g(x))(g_i(x + v) - g_i(x)) + o(g(x + v) - g(x))$$
which by Lemma 19.3.4 equals

\[(f \circ g)(x + v) = f(g(x + v)) = f(g(x)) + \sum_{i=1}^{p} \frac{\partial f(g(x))}{\partial y_i} \left( \sum_{j=1}^{n} \frac{\partial g_i(x)}{\partial x_j} v_j + o(v) \right) + o(v).\]

Now since \(Dg(x)\) exists, the above becomes

\[(f \circ g)(x + v) = f(g(x)) + \sum_{i=1}^{p} \frac{\partial f(g(x))}{\partial y_i} \left( \sum_{j=1}^{n} \frac{\partial g_i(x)}{\partial x_j} v_j + \sum_{i=1}^{p} \frac{\partial f(g(x))}{\partial y_i} o(v) + o(v) \right)\]

because \(\sum_{i=1}^{p} \frac{\partial f(g(x))}{\partial y_i} o(v) + o(v) = o(v)\). This establishes \(19.14\) because of Theorem 19.4.3 on Page 394. Thus

\[\left(D (f \circ g)(x)\right)_{kj} = \sum_{i=1}^{p} \frac{\partial f_k(g(x))}{\partial y_i} \frac{\partial g_i(x)}{\partial x_j} = \sum_{i=1}^{p} Df(g(x))_{ki} Dg(x)_{ij}.\]

Then \(19.16\) follows from the definition of matrix multiplication.

### 19.4 Proof Of The Second Derivative Test

**Definition 19.4.1** The matrix, \(\left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\right)\) is called the Hessian matrix, denoted by \(H(x)\).

Now recall the Taylor formula with the Lagrange form of the remainder. Since most people don’t pay any attention to this important topic when they take calculus, here is a statement and proof of this theorem.

**Theorem 19.4.2** Suppose \(f\) has \(n + 1\) derivatives on an interval, \((a, b)\) and let \(c \in (a, b)\). Then if \(x \in (a, b)\), there exists \(\xi\) between \(c\) and \(x\) such that

\[f(x) = f(c) + \sum_{k=1}^{n+1} \frac{f^{(k)}(c)}{k!} (x - c)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - c)^{n+1}.\]

(In this formula, the symbol \(\sum_{k=1}^{n} a_k \) will denote the number 0.)

**Proof:** If \(n = 0\) then the theorem is true because it is just the mean value theorem. Suppose the theorem is true for \(n - 1, n \geq 1\). It can be assumed \(x \neq c\) because if \(x = c\) there is nothing to show. Then there exists \(K\) such that

\[f(x) - \left(f(c) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!} (x - c)^k + K(x - c)^{n+1}\right) = 0 \quad (19.17)\]
19.4. PROOF OF THE SECOND DERIVATIVE TEST

In fact,

\[ K = -f(x) + \left( f(c) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!} (x-c)^k \right) \]

\[ \frac{1}{(x-c)^{n+1}}. \]

Now define \( F(t) \) for \( t \) in the closed interval determined by \( x \) and \( c \) by

\[ F(t) \equiv f(x) - \left( f(t) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!} (x-t)^k + K(x-t)^{n+1} \right). \]

The \( c \) in \[ ] got replaced by \( t \).

Therefore, \( F(c) = 0 \) by the way \( K \) was chosen and also \( F(x) = 0 \). By the mean value theorem or Rolle’s theorem, there exists \( t_1 \) between \( x \) and \( c \) such that \( F'(t_1) = 0 \). Therefore,

\[ 0 = f'(t_1) - \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!} (x-t_1)^{k-1} - K(n+1)(x-t_1)^n \]

\[ = f'(t_1) - \left( f'(c) + \sum_{k=1}^{n-1} \frac{f^{(k+1)}(c)}{k!} (x-t_1)^k \right) - K(n+1)(x-t_1)^n \]

\[ = f'(t_1) - \left( f'(c) + \sum_{k=1}^{n-1} \frac{f^{(k)}(c)}{k!} (x-t_1)^k \right) - K(n+1)(x-t_1)^n \]

By induction applied to \( f' \), there exists \( \xi \) between \( x \) and \( t_1 \) such that the above simplifies to

\[ 0 = \frac{f^{(n)}(\xi)}{n!} (x-t_1)^n - K(n+1)(x-t_1)^n \]

\[ = \frac{f^{(n+1)}(\xi)}{n!} (x-t_1)^n - K(n+1)(x-t_1)^n \]

therefore,

\[ K = \frac{f^{(n+1)}(\xi)}{(n+1)!} = \frac{f^{(n+1)}(\xi)}{(n+1)!} \]

and the formula is true for \( n \). This proves the theorem.

The term \( \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-c)^{n+1} \), is called the remainder and this particular form of the remainder is called the Lagrange form of the remainder.

Now let \( f : U \to \mathbb{R} \) where \( U \) is an open subset of \( \mathbb{R}^n \). Suppose \( f \in C^2(U) \). Let \( x \in U \) and let \( r > 0 \) be such that \( B(x,r) \subseteq U \).

Then for \( ||v|| < r \) consider

\[ f(x+tv) - f(x) = h(t) \]

for \( t \in [0,1] \). Then from Taylor’s theorem for the case where \( m = 2 \) and the chain rule, using the repeated index summation convention and the chain rule,

\[ h'(t) = \frac{\partial f}{\partial x_i} (x+tv) v_i, \quad h''(t) = \frac{\partial^2 f}{\partial x_i \partial x_j} (x+tv) v_i v_j. \]

Thus

\[ h''(t) = v^T H(x+tv) v. \]
From Theorem 19.4.2 there exists \( t \in (0, 1) \) such that

\[
f(x + v) = f(x) + \frac{\partial f}{\partial x_i}(x) v_i + \frac{1}{2} v^T H(x + tv) v
\]

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

\[
\lim_{|v| \to 0} \frac{1}{2} \frac{(v^T (H(x + tv) - H(x)) v)}{|v|^2} = 0
\]

because of the continuity of the entries of \( H(x) \).

Recall the following important theorem from linear algebra.

**Theorem 19.4.3** 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. In particular, it has all real eigenvalues and an orthonormal basis of eigenvectors.

**Theorem 19.4.4** 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 (19.13) 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 Theorem 19.4.3 \( H(x) \) 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

\[
u^T H(x) u = \left( \sum_{k=1}^n u_k v_k^2 \right) H(x) \left( \sum_{j=1}^n u_j v_j \right)
\]

\[
= \sum_{k,j} u_k v_k^T H(x) v_j u_j
\]

\[
= \sum_{j=1}^n u_j^2 \lambda_j \geq \delta^2 \sum_{j=1}^n u_j^2 = \delta^2 |u|^2.
\]

From 19.20 and 19.13, 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 (19.20), replacing \( v \) with \( sv \) and letting \( t \) depend on \( s \),

\[
f(x+sv) = f(x) + \frac{1}{2} s^2 v^T H(x) v + \frac{1}{2} s^2 (v^T (H(x+sv) - H(x)) v)
\]

which implies

\[
f(x+sv) = f(x) + \frac{1}{2} s^2 \lambda^2 |v|^2 + \frac{1}{2} s^2 (v^T (H(x+sv) - H(x)) v)
\]

\[
\geq f(x) + \frac{1}{4} s^2 \lambda^2 |v|^2
\]

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.
CHAPTER 19. THE DERIVATIVE OF VECTOR VALUED FUNCTIONS, WHAT IS THE DERIVATIVE?
Chapter 20

Implicit Function Theorem*

The implicit function theorem is one of the greatest theorems in mathematics. There are many versions of this theorem which are of far greater generality than the one given here. The proof given here is like one found in one of Caratheodory’s books on the calculus of variations. It is not as elegant as some of the others which are based on a contraction mapping principle but it may be more accessible. However, it is an advanced topic. Don’t waste your time with it unless you have first read and understood the material on rank and determinants found in the chapter on the mathematical theory of determinants. You will also need to use the extreme value theorem for a function of $n$ variables and the chain rule as well as everything about matrix multiplication.

Definition 20.0.5  Suppose $U$ is an open set in $\mathbb{R}^n \times \mathbb{R}^m$ and $(x, y)$ will denote a typical point of $\mathbb{R}^n \times \mathbb{R}^m$ with $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Let $f : U \to \mathbb{R}^p$ be in $C^1(U)$. Then define

$$D_1 f (x, y) = \begin{pmatrix} f_{1, x_1} (x, y) & \cdots & f_{1, x_n} (x, y) \\ \vdots & \ddots & \vdots \\ f_{p, x_1} (x, y) & \cdots & f_{p, x_n} (x, y) \end{pmatrix},$$

$$D_2 f (x, y) = \begin{pmatrix} f_{1, y_1} (x, y) & \cdots & f_{1, y_m} (x, y) \\ \vdots & \ddots & \vdots \\ f_{p, y_1} (x, y) & \cdots & f_{p, y_m} (x, y) \end{pmatrix}.$$  

Thus $D f (x, y)$ is a $p \times (n + m)$ matrix of the form

$$D f (x, y) = \begin{pmatrix} D_1 f (x, y) & D_2 f (x, y) \end{pmatrix}.$$  

Note that $D_1 f (x, y)$ is an $p \times n$ matrix and $D_2 f (x, y)$ is a $p \times m$ matrix.

Theorem 20.0.6 (implicit function theorem) Suppose $U$ is an open set in $\mathbb{R}^n \times \mathbb{R}^m$.

Let $f : U \to \mathbb{R}^n$ be in $C^1(U)$ and suppose

$$f (x_0, y_0) = 0, \ D_1 f (x_0, y_0)^{-1} \text{ exists.}$$  (20.1)
CHAPTER 20. IMPLICIT FUNCTION THEOREM

Then there exist positive constants, $\delta, \eta$, such that for every $y \in B(y_0, \eta)$ there exists a unique $x(y) \in B(x_0, \delta)$ such that

$$f(x(y), y) = 0. \quad (20.2)$$

Furthermore, the mapping, $y \to x(y)$ is in $C^1(B(y_0, \eta))$.

Proof: Let

$$f(x, y) = \begin{pmatrix} f_1(x, y) \\ f_2(x, y) \\ \vdots \\ f_n(x, y) \end{pmatrix}. $$

Define for $(x^1, \ldots, x^n) \in B(x_0, \delta)^n$ and $y \in B(y_0, \eta)$ the following matrix.

$$J(x^1, \ldots, x^n, y) = \begin{pmatrix} f_{1, x^1}(x^1, y) & \cdots & f_{1, x^n}(x^1, y) \\ \vdots & \ddots & \vdots \\ f_{n, x^1}(x^n, y) & \cdots & f_{n, x^n}(x^n, y) \end{pmatrix}. $$

Then by the assumption of continuity of all the partial derivatives and the extreme value theorem, there exists $r > 0$ and $\delta_0, \eta_0 > 0$ such that if $\delta \leq \delta_0$ and $\eta \leq \eta_0$, it follows that for all $(x^1, \ldots, x^n) \in B(x_0, \delta)^n$ and $y \in B(y_0, \eta)$,

$$\det(J(x^1, \ldots, x^n, y)) > r > 0. \quad (20.3)$$

and $B(x_0, \delta_0) \times B(y_0, \eta_0) \subseteq U$. By continuity of all the partial derivatives and the extreme value theorem, it can also be assumed there exists a constant, $K$ such that for all $(x, y) \in B(x_0, \delta_0) \times B(y_0, \eta_0)$ and $i = 1, 2, \cdots, n$, the $i^{th}$ row of $D_2f(x, y)$, given by $D_2f_i(x, y)$ satisfies

$$|D_2f_i(x, y)| < K, \quad (20.4)$$

and for all $(x^1, \ldots, x^n) \in B(x_0, \delta_0)^n$ and $y \in B(y_0, \eta_0)$ the $i^{th}$ row of the matrix,

$$J(x^1, \ldots, x^n, y)^{-1}$$

which equals $e_i^T (J(x^1, \ldots, x^n, y)^{-1})$ satisfies

$$\left|e_i^T (J(x^1, \ldots, x^n, y)^{-1}) \right| < K. \quad (20.5)$$

(Recall that $e_i$ is the column vector consisting of all zeros except for a 1 in the $i^{th}$ position.)

To begin with it is shown that for a given $y \in B(y_0, \eta)$ there is at most one $x \in B(x_0, \delta)$ such that $f(x, y) = 0$.

Pick $y \in B(y_0, \eta)$ and suppose there exist $x, z \in B(x_0, \delta)$ such that $f(x, y) = f(z, y) = 0$. Consider $f_i$ and let

$$h(t) = f_i(x + t(z - x), y).$$

Then $h(1) = h(0)$ and so by the mean value theorem, $h'(t_i) = 0$ for some $t_i \in (0, 1)$. Therefore, from the chain rule and for this value of $t_i$,

$$h'(t_i) = Df_i(x + t_i(z - x), y)(z - x) = 0. \quad (20.6)$$
Then denote by $x^i$ the vector, $x + t_i (z - x)$. It follows from \textbf{20.4} that

$$J (x^1, \ldots, x^n, y) (z - x) = 0$$

and so from \textbf{20.3} $z - x = 0$. (The matrix, in the above is invertible since its determinant is nonzero.) Now it will be shown that if $\eta$ is chosen sufficiently small, then for all $y \in B(y_0, \eta)$, there exists a unique $x (y) \in B(x_0, \delta)$ such that $f (x(y), \eta) = 0$.

**Claim:** If $\eta$ is small enough, then the function, $h_y (x) = |f (x, y)|^2$ achieves its minimum value on $\overline{B(x_0, \delta)}$ at a point of $B(x_0, \delta)$. (The existence of a point in $\overline{B(x_0, \delta)}$ at which $h_y$ achieves its minimum follows from the extreme value theorem.)

**Proof of claim:** Suppose this is not the case. Then there exists a sequence $\eta_k \rightarrow 0$ and for some $y_k$ having $|y_k - y_0| < \eta_k$, the minimum of $h_{y_k}$ on $\overline{B(x_0, \delta)}$ occurs on a point of $\overline{B(x_0, \delta)}$, $x_k$ such that $|x_0 - x_k| = \delta$. Now taking a subsequence, still denoted by $k$, it can be assumed that $x_k \rightarrow x$ with $|x - x_k| = \delta$ and $y_k \rightarrow y_0$. This follows from the fact that $\{x \in \overline{B(x_0, \delta)} : |x - x_k| = \delta\}$ is a closed and bounded set and is therefore sequentially compact. Let $\varepsilon > 0$. Then for $k$ large enough, the continuity of $y \rightarrow h_y (x_0)$ implies $h_{y_k} (x_0) < \varepsilon$ because $h_{y_k} (x_0) = 0$ since $f (x_0, y_0) = 0$. Therefore, from the definition of $x_k$, it is also the case that $h_{y_k} (x_k) < \varepsilon$. Passing to the limit yields $h_{y_k} (x) \leq \varepsilon$. Since $\varepsilon > 0$ is arbitrary, it follows that $h_{y_k} (x) = 0$ which contradicts the first part of the argument in which it was shown that for $y \in B(y_0, \eta)$ there is at most one point, $x \in \overline{B(x_0, \delta)}$ where $f (x, y) = 0$. Here two have been obtained, $x_0$ and $x$. This proves the claim.

Choose $\eta < \eta_0$ and also small enough that the above claim holds and let $x (y)$ denote a point of $B(x_0, \delta)$ at which the minimum of $h_y$ on $\overline{B(x_0, \delta)}$ is achieved. Since $x (y)$ is an interior point, you can consider $h_y (x (y) + tv)$ for $|t|$ small and conclude this function of $t$ has a zero derivative at $t = 0$. Now

$$h_y (x (y) + tv) = \sum_{i=1}^n f_i^2 (x (y) + tv, y)$$

and so from the chain rule,

$$\frac{d}{dt} h_y (x (y) + tv) = \sum_{i=1}^n 2f_i (x (y) + tv, y) \frac{\partial f_i (x (y) + tv, y)}{\partial x_j} v_j.$$ 

Therefore, letting $t = 0$, it is required that for every $v$,

$$\sum_{i=1}^n 2f_i (x (y), y) \frac{\partial f_i (x (y), y)}{\partial x_j} v_j = 0.$$

In terms of matrices this reduces to

$$0 = 2f (x (y), y)^T D_1 f (x (y), y) v$$

for every vector $v$. Therefore,

$$0 = f (x (y), y)^T D_1 f (x (y), y)$$

From \textbf{40.3}, it follows $f (x (y), y) = 0$. This proves the existence of the function $y \rightarrow x (y)$ such that $f (x (y), y) = 0$ for all $y \in B(y_0, \eta)$.

It remains to verify this function is a $C^1$ function. To do this, let $y_1$ and $y_2$ be points of $B(y_0, \eta)$. Then as before, consider the $i^{th}$ component of $f$ and consider the same argument
using the mean value theorem to write

\[ 0 = f_i (x (y_1), y_1) - f_i (x (y_2), y_2) \]
\[ = f_i (x (y_1), y_1) - f_i (x (y_2), y_1) + f_i (x (y_2), y_1) - f_i (x (y_2), y_2) \]
\[ = D_1 f_i (x^i, y_1) (x (y_1) - x (y_2)) + D_2 f_i (x (y_2), y^i) (y_1 - y_2). \]  \hspace{1cm} (20.7)

where \( y^i \) is a point on the line segment joining \( y_1 \) and \( y_2 \). Thus from (20.3) and the Cauchy
Schwarz inequality,

\[ |D_2 f_i (x (y_2), y^i) (y_1 - y_2)| \leq K |y_1 - y_2|. \]

Therefore, letting \( M (y_1, \cdots, y^n) \equiv M \) denote the matrix having the \( i \)th row equal to
\( D_2 f_i (x (y_2), y^i) \), it follows

\[ |M (y_1 - y_2)| \leq \left( \sum_i K^2 |y_1 - y_2|^2 \right)^{1/2} = \sqrt{m} K |y_1 - y_2|. \]  \hspace{1cm} (20.8)

Also, from (20.7),

\[ J (x^1, \cdots, x^n, y_1) (x (y_1) - x (y_2)) = -M (y_1 - y_2) \]  \hspace{1cm} (20.9)

and so from (20.8) and (20.9),

\[ |x (y_1) - x (y_2)| = |J (x^1, \cdots, x^n, y_1)^{-1} M (y_1 - y_2)| \]
\[ = \left( \sum_{i=1}^n |e_i^T J (x^1, \cdots, x^n, y_1)^{-1} M (y_1 - y_2)|^2 \right)^{1/2} \]
\[ \leq \left( \sum_{i=1}^n K^2 |M (y_1 - y_2)|^2 \right)^{1/2} \leq \left( \sum_{i=1}^n K^2 \left( \sqrt{m} K |y_1 - y_2| \right)^2 \right)^{1/2} \]
\[ = K^2 \sqrt{mn} |y_1 - y_2| \]  \hspace{1cm} (20.10)

It follows as in the proof of the chain rule that

\[ o (x (y + v) - x (y)) = o (v). \]  \hspace{1cm} (20.12)

Now let \( y \in B (y_0, \eta) \) and let \( |v| \) be sufficiently small that \( y + v \in B (y_0, \eta) \). Then

\[ 0 = f (x (y + v), y + v) - f (x (y), y) \]
\[ = f (x (y + v), y + v) - f (x (y + v), y) + f (x (y + v), y) - f (x (y), y) \]
\[ = D_2 f (x (y + v), y) v + D_1 f (x (y), y) (x (y + v) - x (y)) + o \left( |x (y + v) - x (y)| \right) \]
\[ = D_2 f (x (y), y) v + D_1 f (x (y), y) (x (y + v) - x (y)) + o \left( |x (y + v) - x (y)| \right) \]
\[ + \left( D_2 f (x (y + v), y) v - D_2 f (x (y), y) v \right) \]
\[ = D_2 f (x (y), y) v + D_1 f (x (y), y) (x (y + v) - x (y)) + o (v). \]

Therefore,

\[ x (y + v) - x (y) = -D_1 f (x (y), y)^{-1} D_2 f (x (y), y) v + o (v) \]
which shows that \( D x (y) = -D_1 f (x (y), y)^{-1} D_2 f (x (y), y) \) and \( y \to D x (y) \) is continuous.
This proves the theorem.
20.1 The Method Of Lagrange Multipliers

As an application of the implicit function theorem, consider the method of Lagrange multipliers. Recall the problem is to maximize or minimize a function subject to equality constraints. Let \( f : U \to \mathbb{R} \) be a \( C^1 \) function where \( U \subseteq \mathbb{R}^n \) and let

\[
g_i(x) = 0, \quad i = 1, \ldots, m
\]

(20.13)

be a collection of equality constraints with \( m < n \). Now consider the system of nonlinear equations

\[
\begin{align*}
f(x) &= a \\
g_i(x) &= 0, \quad i = 1, \ldots, m.
\end{align*}
\]

Recall \( x_0 \) is a local maximum if \( f(x_0) \geq f(x) \) for all \( x \) near \( x_0 \) which also satisfies the constraints \( 20.13 \). A local minimum is defined similarly. Let \( F : U \times \mathbb{R} \to \mathbb{R}^{m+1} \) be defined by

\[
F(x,a) \equiv \begin{pmatrix}
    f(x) - a \\
g_1(x) \\
    \vdots \\
g_m(x)
\end{pmatrix}.
\]

(20.14)

Now consider the \( m + 1 \times n \) matrix,

\[
\begin{pmatrix}
f_{x_1}(x_0) & \cdots & f_{x_n}(x_0) \\
g_{1x_1}(x_0) & \cdots & g_{1x_n}(x_0) \\
\vdots & & \vdots \\
g_{mx_1}(x_0) & \cdots & g_{mx_n}(x_0)
\end{pmatrix}
\]

(20.17)

If this matrix has rank \( m + 1 \) then some \( m + 1 \times m + 1 \) submatrix has nonzero determinant. It follows from the implicit function theorem there exists \( m + 1 \) variables, \( x_1, \ldots, x_{m+1} \) such that the system

\[
F(x,a) = 0
\]

(20.15)

specifies these \( m + 1 \) variables as a function of the remaining \( n - (m + 1) \) variables and \( a \) in an open set of \( \mathbb{R}^{n-m} \). Thus there is a solution \( (x,a) \) to \( 20.15 \) for some \( x \) close to \( x_0 \) whenever \( a \) is in some open interval. Therefore, \( x_0 \) cannot be either a local minimum or a local maximum. It follows that if \( x_0 \) is either a local maximum or a local minimum, then the above matrix must have rank less than \( m + 1 \) which requires the rows to be linearly dependent. Thus, there exist \( m \) scalars,

\( \lambda_1, \ldots, \lambda_m \),

and a scalar \( \mu \), not all zero such that

\[
\mu \begin{pmatrix}
f_{x_1}(x_0) \\
\vdots \\
f_{x_n}(x_0)
\end{pmatrix} = \lambda_1 \begin{pmatrix}
g_{1x_1}(x_0) \\
\vdots \\
g_{1x_n}(x_0)
\end{pmatrix} + \cdots + \lambda_m \begin{pmatrix}
g_{mx_1}(x_0) \\
\vdots \\
g_{mx_n}(x_0)
\end{pmatrix}.
\]

(20.16)

If the column vectors

\[
\begin{pmatrix}
g_{1x_1}(x_0) \\
\vdots \\
g_{1x_n}(x_0)
\end{pmatrix}, \ldots, \begin{pmatrix}
g_{mx_1}(x_0) \\
\vdots \\
g_{mx_n}(x_0)
\end{pmatrix}
\]

(20.17)
are linearly independent, then, \( \mu \neq 0 \) and dividing by \( \mu \) yields an expression of the form
\[
\begin{pmatrix}
  f_{x_1}(x_0) \\
  \vdots \\
  f_{x_n}(x_0)
\end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(x_0) \\
  \vdots \\
  g_{1x_n}(x_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(x_0) \\
  \vdots \\
  g_{mx_n}(x_0) \end{pmatrix}
\]
(20.18)

at every point \( x_0 \) which is either a local maximum or a local minimum. This proves the following theorem.

**Theorem 20.1.1** Let \( U \) be an open subset of \( \mathbb{R}^n \) and let \( f : U \to \mathbb{R} \) be a \( C^1 \) function. Then if \( x_0 \in U \) is either a local maximum or local minimum of \( f \) subject to the constraints \( \mathbf{20.1.1} \), then \( \mathbf{20.1.1} \) must hold for some scalars \( \mu, \lambda_1, \ldots, \lambda_m \) not all equal to zero. If the vectors in \( \mathbf{20.1.4} \) are linearly independent, it follows that an equation of the form \( \mathbf{20.1.8} \) holds.

### 20.2 The Local Structure Of \( C^1 \) Mappings

**Definition 20.2.1** Let \( U \) be an open set in \( \mathbb{R}^n \) and let \( h : U \to \mathbb{R}^n \). Then \( h \) is called primitive if it is of the form
\[
h(x) = \begin{pmatrix} x_1 & \cdots & \alpha(x) & \cdots & x_n \end{pmatrix}^T.
\]

Thus, \( h \) is primitive if it only changes one of the variables. A function, \( F : \mathbb{R}^n \to \mathbb{R}^n \) is called a flip if
\[
F(x_1, \ldots, x_k, \ldots, x_l, \ldots, x_n) = (x_1, \ldots, x_l, \ldots, x_k, \ldots, x_n)^T.
\]

Thus a function is a flip if it interchanges two coordinates. Also, for \( m = 1, 2, \ldots, n \),
\[
P_m(x) \equiv \begin{pmatrix} x_1 & x_2 & \cdots & x_m & 0 & \cdots & 0 \end{pmatrix}^T.
\]

It turns out that if \( h(0) = 0, Dh(0)^{-1} \) exists, and \( h \) is \( C^1 \) on \( U \), then \( h \) can be written as a composition of primitive functions and flips. This is a very interesting application of the inverse function theorem.

**Theorem 20.2.2** Let \( h : U \to \mathbb{R}^n \) be a \( C^1 \) function with \( h(0) = 0, Dh(0)^{-1} \) exists. Then there an open set, \( V \subseteq U \) containing \( 0 \), flips, \( F_1, \ldots, F_{n-1} \), and primitive functions, \( G_n, G_{n-1}, \ldots, G_1 \) such that for \( x \in V \),
\[
h(x) = F_1 \circ \cdots \circ F_{n-1} \circ G_n \circ G_{n-1} \circ \cdots \circ G_1(x).
\]

**Proof:** Let
\[
h_1(x) \equiv h(x) = \begin{pmatrix} \alpha_1(x) & \cdots & \alpha_n(x) \end{pmatrix}^T
\]
\[
Dh(0)e_1 = \begin{pmatrix} \alpha_{1,1}(0) & \cdots & \alpha_{n,1}(0) \end{pmatrix}^T
\]

where \( \alpha_{k,1} \) denotes \( \frac{\partial \alpha_k}{\partial x_1} \). Since \( Dh(0) \) is one to one, the right side of this expression cannot be zero. Hence there exists some \( k \) such that \( \alpha_{k,1}(0) \neq 0 \). Now define
\[
G_1(x) \equiv \begin{pmatrix} \alpha_k(x) & x_2 & \cdots & x_n \end{pmatrix}^T
\]
Then the matrix of \( DG(0) \) is of the form

\[
\begin{pmatrix}
\alpha_{k,1}(0) & \cdots & \cdots & \alpha_{k,n}(0) \\
0 & 1 & 0 \\
\vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1
\end{pmatrix}
\]

and its determinant equals \( \alpha_{k,1}(0) \neq 0 \). Therefore, by the inverse function theorem, there exists an open set, \( U_1 \) containing \( 0 \) and an open set, \( V_2 \) containing \( 0 \) such that \( G_1(U_1) = V_2 \) and \( G_1 \) is one to one and onto such that it and its inverse are both \( C^1 \). Let \( F_1 \) denote the flip which interchanges \( x_k \) with \( x_1 \). Now define

\[
h_2(y) = F_1 \circ h_1 \circ G_1^{-1}(y)
\]

Thus

\[
h_2(G_1(x)) = F_1 \circ h_1(x)
\]

Therefore,

\[
P_1 h_2(G_1(x)) = \begin{pmatrix}
\alpha_k(x) & \cdots & \alpha_1(x) & \cdots & \alpha_n(x)
\end{pmatrix}^T
\]

Also

\[
P_1(G_1(x)) = \begin{pmatrix}
\alpha_k(x) & x_2 & \cdots & x_n
\end{pmatrix}^T
\]

so \( P_1 h_2(y) = P_1(x) \) for all \( y \in V_2 \). Also, \( h_2(0) = 0 \) and \( Dh_2(0)^{-1} \) exists because of the definition of \( h_2 \) above and the chain rule. Also, since \( F_1^2 = \text{identity} \), it follows from (20.20) that

\[
h(x) = h_1(x) = F_1 \circ h_2 \circ G_1(x).
\]

Suppose then that for \( m \geq 2 \),

\[
P_{m-1} h_m(x) = P_{m-1}(x)
\]

for all \( x \in U_m \), an open subset of \( U \) containing \( 0 \) and \( h_m(0) = 0, Dh_m(0)^{-1} \) exists. From (20.21) \( h_m(x) \) must be of the form

\[
h_m(x) = \begin{pmatrix}
x_1 & \cdots & x_{m-1} & \alpha_1(x) & \cdots & \alpha_n(x)
\end{pmatrix}^T
\]

where these \( \alpha_k \) are different than the ones used earlier. Then

\[
Dh_m(0) e_m = \begin{pmatrix}
0 & \cdots & 0 & \alpha_{1,m}(0) & \cdots & \alpha_{n,m}(0)
\end{pmatrix}^T \neq 0
\]

because \( Dh_m(0)^{-1} \) exists. Therefore, there exists a \( k \) such that \( \alpha_{k,m}(0) \neq 0 \), not the same \( k \) as before. Define

\[
G_{m+1}(x) = \begin{pmatrix}
x_1 & \cdots & x_{m-1} & \alpha_k(x) & x_{m+1} & \cdots & x_n
\end{pmatrix}^T
\]

Then \( G_{m+1}(0) = 0 \) and \( DG_{m+1}(0)^{-1} \) exists similar to the above. In fact \( \det(DG_{m+1}(0)) = \alpha_{k,m}(0) \). Therefore, by the inverse function theorem, there exists an open set, \( V_{m+1} \) containing \( 0 \) such that \( V_{m+1} = G_{m+1}(U_m) \) with \( G_{m+1} \) and its inverse being one to one continuous and onto. Let \( F_m \) be the flip which flips \( x_m \) and \( x_k \). Then define \( h_{m+1} \) on \( V_{m+1} \) by

\[
h_{m+1}(y) = F_m \circ h_m \circ G_{m+1}^{-1}(y).
\]
Thus for \( x \in U_m \),

\[
h_{m+1}(G_{m+1}(x)) = (F_m \circ h_m)(x).
\]

(20.23)

and consequently,

\[
F_m \circ h_{m+1} \circ G_{m+1}(x) = h_m(x)
\]

(20.24)

It follows

\[
P_m h_{m+1} (G_{m+1}(x)) = P_m (F_m \circ h_m)(x)
\]

\[
= \begin{pmatrix} x_1 & \cdots & x_{m-1} & \alpha_k(x) & 0 & \cdots & 0 \end{pmatrix}^T
\]

and

\[
P_m (G_{m+1}(x)) = \begin{pmatrix} x_1 & \cdots & x_{m-1} & \alpha_k(x) & 0 & \cdots & 0 \end{pmatrix}^T.
\]

Therefore, for \( y \in V_{m+1} \),

\[
P_m h_{m+1}(y) = P_m(y).
\]

As before, \( h_{m+1}(0) = 0 \) and \( Dh_{m+1}(0)^{-1} \) exists. Therefore, we can apply (20.24) repeatedly, obtaining the following:

\[
h(x) = F_1 \circ h_2 \circ G_1(x)
\]

\[
= F_1 \circ F_2 \circ h_3 \circ G_2 \circ G_1(x)
\]

\[
= \cdots
\]

\[
= F_1 \circ \cdots \circ F_{n-1} \circ h_n \circ G_{n-1} \circ \cdots \circ G_1(x)
\]

where

\[
P_{n-1} h_n(x) = P_{n-1}(x) = \begin{pmatrix} x_1 & \cdots & x_{n-1} & 0 \end{pmatrix}^T
\]

and so \( h_n(x) \) is of the form

\[
h_n(x) = \begin{pmatrix} x_1 & \cdots & x_{n-1} & \alpha(x) \end{pmatrix}^T.
\]

Therefore, define the primitive function, \( G_n(x) \) to equal \( h_n(x) \). This proves the theorem.
Part VI

Multiple Integrals
Chapter 21

The Riemann Integral On $\mathbb{R}^n$

21.1 Methods For Double Integrals

Quiz

1. Maximize $2x + y$ subject to the condition that $\frac{x^2}{4} + \frac{y^2}{9} \leq 1$.

2. 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 $(0, 0, 0)$.

3. Find the points on $y^2x = 9$ which are closest to $(0, 0)$.

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

$$M_Q (f) \equiv \sup \{ f(x) : x \in Q \}, \ m_Q (f) \equiv \inf \{ f(x) : x \in Q \},$$

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 \( \sum_Q m_Q(f) v(Q) \) 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.
21.1. METHODS FOR DOUBLE INTEGRALS

\[(x,y) \in [a,b] \times [c,d],\] means \(x \in [a,b]\) and also \(y \in [c,d]\) and the points which do this comprise the rectangle just as shown in the picture.

**Definition 21.1.1** For \(i = 1,2\), let \(\{\alpha_k^i\}_{k=-\infty}^{\infty}\) be points on \(\mathbb{R}\) which satisfy

\[
\lim_{k \to \infty} \alpha_k^i = \infty, \quad \lim_{k \to -\infty} \alpha_k^i = -\infty, \quad \alpha_k^i < \alpha_k^{i+1}.
\]

(21.2)

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

\[
Q = [\alpha_k^1, \alpha_{k+1}^1] \times [\alpha_l^2, \alpha_{l+1}^2].
\]

(21.3)

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,

\[
\sum_{Q \in \mathcal{G}} m_Q(f) v(Q)
\]

is called the 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 21.1.2** 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

$$\int_{c}^{d} \left( \int_{a}^{b} f(x,y) \, dx \right) \, dy \tag{21.4}$$

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,

$$\int_{a}^{b} \left( \int_{c}^{d} f(x,y) \, dy \right) \, dx \tag{21.5}$$

In terms of the little rectangles, when you take $f(x,y) \, dx$, and add these together using the integral to get

$$\int_{a}^{b} f(x,y) \, dx$$

and then multiply by $dy$ to get

$$\int_{a}^{b} f(x,y) \, dx \, dy$$

you are essentially adding up the contributions to the integral for a fixed $y$. Then you add all these together to get all contributions yielding

$$\int_{c}^{d} \int_{a}^{b} f(x,y) \, dx \, dy$$

Imagine a kind of distributive law holding in a formal sense.
These expressions in 21.4 and 21.5 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 $R^2$. For example, suppose $f$ is defined on the set, $S \subseteq R^2$. What is meant by $\int_S f \, dV$?

**Definition 21.1.3** Let $f : S \to R$ where $S$ is a subset of $R^2$. 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 21.1.4** 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.

$$\int_0^2 \int_0^1 (x^2y + yx) \, dx \, dy = \int_0^1 (x^2y + yx) \, dx = \frac{5}{6}.$$

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 21.1.5 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 \( \int_R f \, dV \). Note that to find the area of this region, you simply integrate the function 1.

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 21.1.6 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 \( \int_R f \, dV \).
21.1. METHODS FOR DOUBLE INTEGRALS

Put the integral with respect to $x$ on the outside first. Then

$$
\int_R f \, dV = \int_0^4 \int_0^{2x} (x^2y) \, dy \, dx
$$

because for each $x \in [0, 4]$, $y$ goes from 0 to $2x$. Then

$$
\int_0^{2x} (x^2y) \, dy = 2x^4
$$

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

$$
\int_0^8 \int_{y/2}^4 (x^2y) \, dx \, dy.
$$

Now

$$
\int_{y/2}^4 (x^2y) \, dx = \frac{64}{3}y - \frac{1}{24}y^4
$$

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 21.1.7** For $R$ as shown below, find $\int_R \sin(y^2) \, dV$.

![Diagram of R](image)

Setting this up to have the integral with respect to $y$ on the inside yields

$$
\int_0^4 \int_{2x}^8 \sin(y^2) \, dy \, dx.
$$
Unfortunately, there is no antiderivative in terms of elementary functions for \( \sin(y^2) \) 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.

### 21.1.1 Density Mass And Center Of 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 is
\[
\int_U \rho \, dV
\]
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 21.1.8** 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 region R](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.
21.2 DOUBLE INTEGRALS IN POLAR COORDINATES

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.

The concept of center of mass of a plate occupying the bounded open set, \( U \) is also easy to express in terms of double integrals. Letting \( \rho \) denote the density of the plate, the moment of a small chunk of mass having coordinates \((x, y)\) about the \( y \) axis is \( x\rho(x, y) \, dV \) and the moment of the same small chunk of mass about the \( x \) axis is \( y\rho(x, y) \, dV \). Therefore the center of mass, \((x_c, y_c)\) is defined in the usual way.

Definition 21.1.9 The center of mass of a plate occupying the bounded open set, \( U \) is defined as \((x_c, y_c)\) where

\[
x_c = \frac{\int_U x\rho(x, y) \, dV}{\int_U \rho(x, y) \, dV}, \quad y_c = \frac{\int_U y\rho(x, y) \, dV}{\int_U \rho(x, y) \, dV}.
\]

In other words, the total moment about the \( y \) axis equals \( x_c \) times the total mass. That is, if you placed the total mass at the single point, \((x_c, y_c)\) this point mass would produce the same moments about the \( x \) and \( y \) axes as the original plate.

Example 21.1.10 In Example 21.1.8, suppose the density is \( \rho(x, y) = y \) as it is in that example. Find the total mass and the center of mass.

First, the total mass was found above. Then the center of mass is

\[
x_c = \frac{\int_0^3 \int_{\sqrt{x/3}}^{\sqrt{x/9}} xy \, dy \, dx \; + \; \int_{\sqrt{2}}^{2} \int_{\sqrt{x/9}}^{\sqrt{2-x}} xy \, dy \, dx}{\int_0^3 \int_{\sqrt{x/9}}^{\sqrt{x/3}} y \, dy \, dx \; + \; \int_{\sqrt{2}}^{2} \int_{\sqrt{x/9}}^{\sqrt{2-x}} y \, dy \, dx} = \frac{39}{16} = 39 \quad \frac{1}{16}
\]

\[
y_c = \frac{\int_0^3 \int_{\sqrt{x/9}}^{\sqrt{x/3}} y^2 \, dy \, dx \; + \; \int_{\sqrt{2}}^{2} \int_{\sqrt{x/9}}^{\sqrt{2-x}} y^2 \, dy \, dx}{\int_0^3 \int_{\sqrt{x/9}}^{\sqrt{x/3}} y \, dy \, dx \; + \; \int_{\sqrt{2}}^{2} \int_{\sqrt{x/9}}^{\sqrt{2-x}} y \, dy \, dx} = \frac{47}{80}
\]

Thus the center of mass is \( \left( \frac{39}{16}, \frac{47}{80} \right) \).

21.2 Double Integrals In Polar Coordinates

Remember polar coordinates,

\[
x = r \cos \theta \]
\[
y = r \sin \theta
\]

where \( \theta \in [0, 2\pi] \) and \( r > 0 \). If you assign a given value to \( r \), the points obtained yield a circle and if you give a value to \( \theta \) the points yield a ray from the origin. Thus assigning many different values for \( r \) and many different values for \( \theta \) yields a grid of the sort illustrated in the following picture.
By contrast, the grid on the right is obtained by assigning different values for \( x \) and \( y \). For the grid on the right, if the vertical lines are \( dx \) apart and the horizontal lines are \( dy \) apart, the area of one of those little boxes would be \( dx \, dy \). This is the increment of area in rectangular coordinates. Now consider the grid on the left which is obtained by setting each of the two polar variables equal to various constants. What is the area of one of those little curvy rectangles if the values for \( r \) and \( \theta \) are very small? Zoom in on one of them as illustrated in the following picture.

The angle between the two straight lines is \( d\theta \) and so the length of one of the curved sides is approximately \( rd\theta \) while the length of the straight sides is \( dr \). Therefore, the area of the little curvy rectangle is approximately equal to \( rdrd\theta \). This is the increment of area in polar coordinates.

Later, I will present a unified way to change variables. For now, consider the following problems which illustrate the use of polar coordinates to compute integrals over areas.

**Example 21.2.1** Find the area of a circle of radius \( R \).

Denote by \( D \) this circle. Then the area of the circle is \( \int dA \) and you need to write \( dA \) in terms of polar coordinates. As described above, \( dA = rdrd\theta \). To compute the integral, note that in terms of the variables, \( \theta \) and \( r \), this region is actually the rectangle, \([0, R] \times [0, 2\pi]\). Therefore, the integral equals

\[
\int_0^{2\pi} \int_0^R rdrd\theta = \pi R^2
\]

which you have already heard about.

**Example 21.2.2** Find the volume of the ball of radius \( R \).

It is enough to find the volume of the top half of this ball and then multiply it by 2. Corresponding to the small curvy rectangle as described above having polar coordinates \((r, \theta)\) the height of the ball over this point is \( \sqrt{R^2 - r^2} \). Therefore, the volume of a small prism having as a base the small curvy rectangle described above is \( \sqrt{R^2 - r^2} rdrd\theta \). Summing these using the integral, the desired volume is

\[
\int_D \sqrt{R^2 - r^2} dA = \int_0^{2\pi} \int_0^R \sqrt{R^2 - r^2} rdrd\theta = \frac{2}{3} \pi R^3
\]
and so the volume of the whole ball is 

\[ \frac{4}{3} \pi R^3 \]

which is another formula you might have seen.

**Example 21.2.3** Find the area inside \( r = 1 + \cos \theta \) for \( \theta \in [0, 2\pi] \).

This is the graph of a cardioid. You saw this in beginning calculus. Let its inside be denoted by \( C \) for cardioid. Then the desired area is

\[ \int_C dA \]

and you need to set up an iterated integral and put in the correct form for \( dA \). For each \( \theta \), you have that \( r \) goes from 0 to \( 1 + \cos \theta \). Therefore, the desired area is given by the iterated integral,

\[ \int_0^{2\pi} \int_0^{1+\cos \theta} r \, dr \, d\theta = \frac{3}{2} \pi \]

This was really easy because of polar coordinates. If you try to do this in rectangular coordinates it will not work very well.

**21.3 A Few Applications**

**Example 21.3.1** A plate occupies the region inside the curve, \( r = 2 \cos \theta \) for \( \theta \in [-\frac{\pi}{2}, \frac{\pi}{2}] \). The density in terms of polar coordinates is \( \delta = r \). Find the mass and center of mass of this plate.

First of all the mass is given by

\[ \int_{-\pi/2}^{\pi/2} \int_{-\pi/2}^{2 \cos \theta} r^2 \, dr \, d\theta = \frac{32}{9} \]

The volume element is \( r \, dr \, d\theta \) and I summed these up multiplied, by the density which was \( r \) and that is the above integral.

Now to compute the center of mass, recall

\[ x_c = \frac{\int_U x \rho(x,y) \, dV}{\int_U \rho(x,y) \, dV}, \quad y_c = \frac{\int_U y \rho(x,y) \, dV}{\int_U \rho(x,y) \, dV} \]

I need to place \( x \) and \( y \) in terms of the polar coordinates. Thus \( x = r \cos \theta, y = r \sin \theta \). A small contribution to the moment about the \( y \) axis is \( r \cos \theta \times r \times r \, dr \, d\theta \). Thus

\[ x_c = \frac{\int_{-\pi/2}^{\pi/2} \int_0^{2 \cos \theta} r^2 \cos \theta \, r \, dr \, d\theta}{\int_{-\pi/2}^{\pi/2} \int_0^{2 \cos \theta} r \, dr \, d\theta} = \frac{6}{5} \]

Similarly,

\[ y_c = \frac{\int_{-\pi/2}^{\pi/2} \int_0^{2 \cos \theta} r^2 \sin \theta \, r \, dr \, d\theta}{\int_{-\pi/2}^{\pi/2} \int_0^{2 \cos \theta} r^2 \, dr \, d\theta} = 0 \]
Example 21.3.2 Let a plate occupy the region, \( C \) which is inside the polar graph, \( r = 2 + \cos \theta \) for \( \theta \in [0, 2\pi] \). Suppose the density of this plate is given by \( \delta (r, \theta) = r \). Find the mass and center of mass of the plate.

Here you need to evaluate the following to get the total mass.

\[
\int_0^{2\pi} \int_0^{2+\cos \theta} r \times r \, dr \, d\theta = \int_0^{2\pi} \int_0^{2+\cos \theta} r^2 \, dr \, d\theta = \frac{22}{3} \pi
\]

Now recall the center of mass is given by

\[
x_c = \frac{\int_U x \rho (x,y) \, dV}{\int_U \rho (x,y) \, dV}, \quad y_c = \frac{\int_U y \rho (x,y) \, dV}{\int_U \rho (x,y) \, dV}
\]

Thus

\[
x_c = \frac{\int_0^{2\pi} \int_0^{2+\cos \theta} (r \cos \theta) \, r^2 \, dr \, d\theta}{\frac{22}{3} \pi} = \frac{57}{44}
\]

and

\[
y_c = \frac{\int_0^{2\pi} \int_0^{2+\cos \theta} (r \sin \theta) \, r^2 \, dr \, d\theta}{\frac{22}{3} \pi} = 0
\]

I think this might be impossible if you tried to do it in rectangular coordinates. However, it is just a little tedious in polar coordinates. Be sure you understand the set up. This is usually the thing which gives people the most trouble in these kinds of problems.

Example 21.3.3 Let \( f (x,y) = \sin (x^2 + y^2) \) for \( (x,y) \) in the circle, \( D = \{(x,y) : x^2 + y^2 \leq 9\} \).

Find

\[
\int_D f \, dA.
\]

You don’t want to try this in rectangular coordinates even though the function is given in rectangular coordinates. You should change it to polar coordinates for two reasons. The first is that \( x^2 + y^2 = r^2 \) and it is easier to look at \( \sin (r^2) \) than \( \sin (x^2 + y^2) \). The main reason is that the integration is taking place on a circle which is a rectangle in polar coordinates and as explained earlier, it is easy to integrate over rectangles. In this case the rectangle is \([0, 2\pi] \times [0, 3]\). Thus the integral to work is

\[
\int_0^{2\pi} \int_0^{3} \sin (r^2) \, r \, dr \, d\theta = -\pi \cos 9 + \pi.
\]

Example 21.3.4 Remember the formula for the area between two polar graphs \( r = f (\theta) \) and \( r = g (\theta) \), \( g (\theta) > f (\theta) \) for \( \theta \in [a,b] \) is given by

\[
\frac{1}{2} \int_a^b (g (\theta)^2 - f (\theta)^2) \, d\theta.
\]

Show this formula from one variable calculus follows from the form of the area increment given here.

Denote by \( R \) the region between the two graphs. Then you need to find

\[
\int_R dA = \int_a^b \int_{f(\theta)}^{g(\theta)} r \, dr \, d\theta = \frac{1}{2} \int_a^b (g (\theta)^2 - f (\theta)^2) \, d\theta \quad (21.6)
\]

which is the formula done earlier.

Here is an example.
Example 21.3.5 Find the area of the region inside the cardioid, \( r = 1 + \cos \theta \) and outside the circle, \( r = 1 \) for \( \theta \in \left[ -\frac{\pi}{2}, \frac{\pi}{2} \right] \).

As is usual in such cases, it is a good idea to graph the curves involved to get an idea what is wanted. It is very important to figure out which function is farther from the origin.

Then you need
\[
\int_{-\pi/2}^{\pi/2} \int_1^{1+\cos \theta} r dr d\theta = 2 + \frac{1}{4} \pi
\]
You could also work it using the formula derived in 21.6 which is like what you did in one variable calculus.

Example 21.3.6 Let \( f(x,y) = e^{x^2+y^2} \) for \((x,y)\) in the pie shaped region \( P \) defined by \( r \in [0,2] \) and \( \theta \in [0, \pi/6] \).

Be sure you can see why the integral wanted is
\[
\int_0^2 \int_0^{\pi/6} e^{r^2} r dr d\theta = \frac{1}{12} \pi - \frac{1}{12} \pi
\]

Example 21.3.7 Find
\[
\int_{-1}^{1} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \sqrt{1+x^2+y^2} dy dx.
\]

In this example you are integrating the function, \( f(x,y) = \sqrt{1+x^2+y^2} \) over the circle of radius 1 centered at the origin. Therefore, changing to polar coordinates it equals
\[
\int_0^{2\pi} \int_0^1 \sqrt{1+r^2} r dr d\theta = \frac{4}{3} \pi \sqrt{2} - \frac{2}{3} \pi
\]
In this case, I think you could have done it without changing to polar coordinates but it would involve wading through much affliction and sorrow. Of course if you like adversity, you could try to do it this way.

Example 21.3.8 Find \( \int_0^\infty e^{-x^2} dx \).

Let \( I = \int_0^\infty e^{-x^2} dx \). Then
\[
I^2 = \left( \int_0^\infty e^{-x^2} dx \right) \left( \int_0^\infty e^{-y^2} dy \right) = \int_0^\infty \int_0^\infty e^{-(x^2+y^2)} dxdy = \int_0^{\pi/2} \int_0^\infty e^{-r^2} r dr d\theta = \frac{1}{4} \pi
\]
CHAPTER 21. THE RIEMANN INTEGRAL ON $\mathbb{R}^N$

It follows $I = \sqrt{\pi}$. This is a very important formula. You showed, (hopefully) in one variable calculus that this integral exists. Now with the aid of polar coordinates you can actually find it.

Consider a rigid plate $P$ in the plane. It is rotating about either the $x$ axis, $y$ axis, or $z$ axis with angular speed $\omega$. This means that at a distance $r$ from the axis of rotation, the speed of a point is $r\omega$. Consider the kinetic energy of this object. Imagine an infinitesimal chunk of area at a point of the plate $dA$. Then the mass of this chunk would equal $\rho dA$ where $\rho$ is the density. It follows that if $r$ is the distance to the axis of rotation, the kinetic energy of this little chunk of mass is

$$\frac{1}{2} \rho r^2 \omega^2 dA$$

Then the kinetic energy of this rotating rigid body is seen to be

$$KE = \int_P \frac{1}{2} \rho r^2 \omega^2 dA = \frac{1}{2} \omega^2 \int_P \rho r^2 dA$$

You can factor out the $\omega^2$ because the plate is rigid. Thus all points have the same angular velocity. The expression in the integral is called the moment of inertia. If the rotation is about the $x$ axis, it is called $I_x$, if the rotation is about the $y$ axis, it is called $I_y$ and if the rotation is about the $z$ axis, it is called either $I_z$ or $I_0$. I have seen it called either one of these. You see that the distance to the axis of rotation is a function of position and so it is changing depending on the point of $P$. This is why an integral is needed. Then denoting this moment of inertia as $I$,

$$KE = \frac{1}{2} I \omega^2$$

so in terms of rotating bodies, $I$ acts like the mass in the usual form of kinetic energy and $\omega$ acts like the speed.

Example 21.3.9 Let $P$ be a plate which occupies the interior of the cardioid $r = 1 + \cos \theta$ and let the density be $r = \rho$. Find $I_z$, $I_x$, $I_y$.

First, let's find $I_z$. From the definition, this is nothing more than

$$\int_P y^2 r dA = \int_0^{2\pi} \int_0^{1+\cos \theta} (r \sin \theta)^2 r r dr d\theta = \frac{63}{20} \pi$$

Now $I_x$ involves rotation about the $x$ axis so here you would want

$$\int_P y^2 r dA = \int_0^{2\pi} \int_0^{1+\cos \theta} (r \sin \theta)^2 r r dr d\theta = \frac{33}{40} \pi$$

Next $I_x$ equals

$$\int_P x^2 r dA = \int_0^{2\pi} \int_0^{1+\cos \theta} (r \cos \theta)^2 r r dr d\theta = \frac{93}{40} \pi$$

21.4 Surface Area And Integrals

The case where a surface is of the form $z = f(x, y)$ for $(x, y)$ in a region $R$ is discussed here. In the following picture you have a surface and $(x, y)$ is found in $R$ which is in the $xy$ plane.
Two curves in the surface are

\[ x \to \begin{pmatrix} x \\ y \\ f(x, y) \end{pmatrix}, \quad y \to \begin{pmatrix} x \\ y \\ f(x, y) \end{pmatrix} \]

Then, as shown in the picture the area of a little section of the surface would be approx-
imately equal to

\[
dS = \left| \begin{pmatrix} 1 & 0 \\ f_x(x,y) & f_y(x,y) \end{pmatrix} \right| dx \times \left| \begin{pmatrix} 0 & 1 \\ f_y(x,y) & f_y(x,y) \end{pmatrix} \right| dy
\]

\[
= \left| \begin{pmatrix} 1 & 0 \\ f_x(x,y) & f_y(x,y) \end{pmatrix} \right| dxdy
\]

\[
= \sqrt{1 + f_x^2(x,y) + f_y^2(x,y)}dxdy
\]

and this is the increment of area on the surface \(z = f(x,y)\).

It might help to think of a lizard. The infinitesimal parallelogram is 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 If the scales are small enough, their sum would serve as a good approximation to the area of the lizard.\(^\text{1}\)

There is also a simple geometric description of these area increments which is a little less formal. 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 \(\frac{1}{\sqrt{1 + f_x^2 + f_y^2}}(-f_x, -f_y, 1)\). Now consider the following picture.

\[\text{[Image of a lizard]}
\]

\[\text{[Diagram showing the angle and normal of a surface element]}
\]

\(^1\)This beautiful lizard is a *Sceloporus magister*. It was photographed by C. Riley Nelson who is in the Zoology department at Brigham Young University © 2004 in Kane Co. Utah. The lizard is a little less than one foot in length.
21.4. SURFACE AREA AND INTEGRALS

In this picture, you are looking at a chunk of area $dS$ on the surface seen on edge and so it seems reasonable to expect to have $dA = dS \cos \theta$. 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,

$$dS = \frac{dA}{\cos (\theta)} = \frac{dA}{\left(1/\sqrt{1 + f_x^2 + f_y^2}\right)} = \sqrt{1 + f_x^2 + f_y^2} dA$$

as claimed.

Now that we know the increment of area on the surface, it is easy to define the integral. This is philosophically the same as before.

**Definition 21.4.1** For $S$ such a surface and $g$ a function defined on $S$, (thus $S$ consists of points $(x, y, f(x, y))$ where $(x, y) \in \mathbb{R}$.) we define

$$\int_S g dS = \int_R g(x, y, f(x, y)) \sqrt{1 + f_x^2(x, y) + f_y^2(x, y)} dA$$

It has been presented in terms of rectangular coordinates so the description of $dA$ will be $dxdy$. Or course, you might want to change the variables to polar coordinates if it would make the evaluation of that integral on the right easier.

**Example 21.4.2** Let $z = \sqrt{x^2 + y^2}$ where $(x, y) \in U$ for $U = \{(x, y) : x^2 + y^2 \leq 4\}$ Find

$$\int_S h dS$$

where $h(x, y, z) = x + z$ and $S$ is the surface described as $(x, y, \sqrt{x^2 + y^2})$ for $(x, y) \in U$.

You want to find

$$\int_U \left(x + \sqrt{x^2 + y^2}\right) \sqrt{1 + \left(\frac{x}{\sqrt{x^2 + y^2}}\right)^2 + \left(\frac{y}{\sqrt{x^2 + y^2}}\right)^2} dxdy$$

Of course it would be much easier to change this to polar coordinates. You notice that $U$ is just a disk of radius 2. Then this is

$$\int_0^{2\pi} \int_0^2 (r \cos (\theta) + r) \sqrt{1 + r^2} r dr d\theta = \frac{16}{3} \sqrt{2\pi}$$

**Example 21.4.3** Let $z = x^2 + y^2$ for $(x, y)$ inside the pie shaped region whose polar coordinates are $r, \theta$ where $r \leq 2$ and $\theta \in [0, \pi/6]$. Find the area of this surface.

You just set up the integral. $\int_P \sqrt{1 + 4x^2 + 4y^2} dA$. Now you want to change to polar coordinates.

$$\int_0^{\pi/6} \int_0^2 \sqrt{1 + 4r^2} r dr d\theta = \frac{17}{72} \sqrt{17\pi} - \frac{1}{72} \pi$$
21.5 Methods For Triple Integrals

21.5.1 Definition Of The Integral

The integral of a function of three variables is similar to the integral of a function of two variables.

Definition 21.5.1 For $i = 1, 2, 3$ let $\{\alpha^i_k\}_{k=-\infty}^{\infty}$ be points on $\mathbb{R}$ which satisfy

$$
\lim_{k \to \infty} \alpha^i_k = \infty, \quad \lim_{k \to -\infty} \alpha^i_k = -\infty, \quad \alpha^i_k < \alpha^i_{k+1}.
$$

(21.7)

For such sequences, define a grid on $\mathbb{R}^3$ denoted by $G$ or $F$ as the collection of boxes of the form

$$
Q = [\alpha^1_k, \alpha^1_{k+1}] \times [\alpha^2_l, \alpha^2_{l+1}] \times [\alpha^3_p, \alpha^3_{p+1}].
$$

(21.8)

If $G$ is a grid, $F$ is called a refinement of $G$ if every box of $G$ is the union of boxes of $F$.

For $G$ a grid,

$$
\sum_{Q \in G} M_Q(f) v(Q)
$$

is the upper sum associated with the grid, $G$ where

$$
M_Q(f) \equiv \sup \{ f(x) : x \in Q \}
$$

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

$$
m_Q(f) \equiv \inf \{ f(x) : x \in Q \}
$$

the lower sum associated with this partition is

$$
\sum_{Q \in G} m_Q(f) v(Q),
$$

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 21.5.2 Let $f : \mathbb{R}^3 \to \mathbb{R}$ be a bounded function which equals zero outside 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, $dV$. The integral involves taking this little 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.

Integrals of functions of three variables are also evaluated by using iterated integrals. Imagine a sum of the form $\sum_{ijk} a_{ijk}$ where there are only finitely many choices for $i, j,$ and $k$ 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.

Having discussed double and triple integrals, the definition of the integral of a function of \( n \) variables is accomplished in the same way.

**Definition 21.5.3** For \( i = 1, \cdots, n \), let \( \{ \alpha_i^k \}_{k = -\infty}^{\infty} \) be points on \( \mathbb{R} \) which satisfy
\[ \lim_{k \to \infty} \alpha_i^k = \infty, \quad \lim_{k \to -\infty} \alpha_i^k = -\infty, \quad \alpha_i^k < \alpha_i^{k+1}. \quad (21.9) \]
For such sequences, define a grid on \( \mathbb{R}^n \) denoted by \( G \) or \( F \) as the collection of boxes of the form
\[ Q = \prod_{i=1}^{n} [a_i, b_i]. \quad (21.10) \]
If \( G \) is a grid, \( F \) is called a refinement of \( G \) if every box of \( G \) is the union of boxes of \( F \).

**Definition 21.5.4** Let \( f \) be a bounded function which equals zero off a bounded set, \( D \), and let \( G \) be a grid. For \( Q \in G \), define
\[ M_Q(f) \equiv \sup \{ f(x) : x \in Q \}, \quad m_Q(f) \equiv \inf \{ f(x) : x \in Q \}. \quad (21.11) \]
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, \( U_G(f) \) 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 21.5.5** For \( E \subseteq \mathbb{R}^n \),
\[ \chi_E(x) \equiv \begin{cases} 1 & \text{if } x \in E \\ 0 & \text{if } x \notin E \end{cases}. \]
Define \( \int_E f \, dV \equiv \int \chi_E f \, dV \) when \( f \chi_E \in \mathcal{R}(\mathbb{R}^n) \).

\(^2\)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.
21.5.2 Iterated Integrals

As before, the integral is often computed by using an iterated integral. In general it is impossible to set up an iterated integral for finding \( \int_E f \, dV \) for arbitrary regions, \( E \) but when the region is sufficiently simple, one can make progress. Suppose the region, \( E \) over which the integral is to be taken is of the form 
\[
E = \{ (x,y,z) : a(x,y) \leq z \leq b(x,y) \}
\]
for \((x,y) \in \mathbb{R}^2\), a two dimensional region. This is illustrated in the following picture in which the bottom surface is the graph of \( z = a(x,y) \) and the top is the graph of \( z = b(x,y) \).

\begin{center}
\includegraphics[width=0.5\textwidth]{image}
\end{center}

Then
\[
\int_E f \, dV = \int_R \left( \int_{a(x,y)}^{b(x,y)} f(x,y,z) \, dz \right) \, dA
\]

Why does this make sense? You pick an \((x,y)\) in \( R \) and then you have \( f(x,y,z) \, dz \) is a small contribution to that inside integral. That inside integral says to add these together. Then you multiply this by \( dA \). This is analogous to taking the value of \( f \) at each of the boxes in the column of the above picture, multiplying by the volume of the box, and adding these together. Next you add over all of \( R \). This would formally add all the \( f(x,y,z) \, dz \, dA \) together where it is helpful to think of \( dV = dz \, dA \).

Now \( \int_{a(x,y)}^{b(x,y)} f(x,y,z) \, dz \) is a function of \( x \) and \( y \) and so you have reduced the triple integral to a double integral over \( R \) of this function of \( x \) and \( y \). Similar reasoning would apply if the region in \( \mathbb{R}^3 \) were of the form
\[
\{ (x,y,z) : a(y,z) \leq x \leq b(y,z) \}
\]
or
\[
\{ (x,y,z) : a(x,z) \leq y \leq b(x,z) \}.
\]

Example 21.5.6 Find the volume of the region, \( E \) in the first octant between \( z = 1-(x+y) \) and \( z = 0 \).

In this case, \( R \) is the region shown.
Thus the region, $E$ is between the plane $z = 1 - (x + y)$ on the top, $z = 0$ on the bottom, and over $R$ shown above. Thus

$$
\int_E 1\,dV = \int_R \int_0^{1-(x+y)} dz\,dA = \int_0^1 \int_0^{1-x} \int_0^{1-(x+y)} dz\,dy\,dx = \frac{1}{6}
$$

Of course iterated integrals have a life of their own although this will not be explored here. You can just write them down and go to work on them. Here are some examples.

**Example 21.5.7** Find $\int_2^3 \int_3^x \int_{3y}^x (x - y) \,dz\,dy\,dx$.

The inside integral yields $\int_{3y}^x (x - y) \,dz = x^2 - 4xy + 3y^2$. Next this must be integrated with respect to $y$ to give $\int_3^x \left( x^2 - 4xy + 3y^2 \right) \,dy = -3x^2 + 18x - 27$. Finally the third integral gives

$$
\int_2^3 \int_3^x \int_{3y}^x (x - y) \,dz\,dy\,dx = \int_2^3 \left( -3x^2 + 18x - 27 \right) \,dx = -1.
$$

**Example 21.5.8** Find $\int_0^\pi \int_0^y \int_0^{y+z} \cos (x + y) \,dx\,dz\,dy$.

The inside integral is $\int_0^{y+z} \cos (x + y) \,dx = 2 \cos z \sin y \cos y + 2 \sin z \cos^2 y - \sin y$. Now this has to be integrated.

$$
\int_0^3 \int_0^{y+z} \cos (x + y) \,dx \,dz = \int_0^3 y \left( 2 \cos z \sin y \cos y + 2 \sin z \cos^2 y - \sin y \right) \,dz
$$

$$
= -1 - 16 \cos^5 y + 20 \cos^3 y - 5 \cos y - 3 (\sin y) y + 2 \cos^2 y.
$$

Finally, this last expression must be integrated from $0$ to $\pi$. Thus

$$
\int_0^\pi \int_0^3 \int_0^{y+z} \cos (x + y) \,dx \,dz \,dy
$$

$$
= \int_0^\pi \left( -1 - 16 \cos^5 y + 20 \cos^3 y - 5 \cos y - 3 (\sin y) y + 2 \cos^2 y \right) \,dy
$$

$$
= -3\pi.
$$

**Example 21.5.9** Here is an iterated integral: $\int_0^2 \int_0^{3-x} \int_0^x dz\,dy\,dx$. 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
\[
\int_0^2 \int_0^{3-\frac{3}{2}x} f(x,y) \, dy \, dx.
\]
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 $f(x,y)$,
\[
\int_0^3 \int_0^{\frac{3}{2}(3-y)} x^2 \, dx \, dy.
\]

**Example 21.5.10** 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 first of the following pictures corresponds to $x = 0$.

Therefore, the outside integrals taken with respect to $z$ and $y$ are of the form $\int_0^\frac{2}{3} \int_{\sqrt{9-y^2}}^{-\sqrt{9-y^2}} dz \, dy$ 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
\[
\int_0^\frac{2}{3} \int_{\sqrt{9-y^2}}^{-\sqrt{9-y^2}} f(x,y) dx \, dy = \frac{860}{243}.
\]

**Example 21.5.11** 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
\[
\int_{-3}^{3} \int_{-\sqrt{9-y^2}}^{\sqrt{9-y^2}} \int_{-\sqrt{9-y^2}}^{\sqrt{9-y^2}} dz \, dx \, dy = 144.
\]
21.5. METHODS FOR TRIPLE INTEGRALS

Note I drew no picture of the three dimensional region. If you are interested, here it is. The two are viewed from different locations.

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.

21.5.3 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 21.5.12 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.

\[
\begin{array}{c}
1 \\
5
\end{array}
\]

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 $z = 5 - (\frac{1}{5}x + y)$. Therefore, an iterated integral which yields the volume is

$$\int_0^5 \int_{1-\frac{1}{5}x}^{5-5y} dz \, dy \, dx = \frac{25}{6}.$$  

Example 21.5.13 Find the mass of the bounded region, $R$ formed by the plane $\frac{1}{3}x + \frac{1}{3}y + \frac{1}{3}z = 1$ 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
\[
\int_0^3 \int_0^{\sqrt{9-x^2}} \int_0^{\frac{9}{4}x - \frac{3}{4}y} z \, dz \, dy \, dx = \frac{75}{8}.
\]

**Example 21.5.14** 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 21.5.15** Find the volume of the bounded region determined by \( x \geq 0, y \geq 0, z \geq 0 \), and \( \frac{1}{7}x + \frac{1}{4}y + \frac{1}{4}z = 1 \), and \( x + \frac{1}{4}y + \frac{1}{4}z = 1 \).

When \( z = 0 \), the plane \( \frac{1}{7}x + \frac{1}{4}y + \frac{1}{4}z = 1 \) intersects the \( xy \) plane in the line whose equation is
\[ \frac{1}{7}x + y = 1 \]
while the plane, \( x + \frac{1}{4}y + \frac{1}{4}z = 1 \) 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 \( 4 \left(1 - \frac{1}{4}x - y\right) \). Similarly, on \( R_2 \), \( z \) goes from 0 to \( 4 \left(1 - \frac{1}{4}y - x\right) \). Therefore, the integral needed is
\[
\int_{R_1} \int_0^{4(1 - \frac{1}{4}x - y)} dz \, dV + \int_{R_2} \int_0^{4(1 - \frac{1}{4}y - x)} dz \, dV
\]
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 \( \left( \frac{7}{8}, \frac{7}{8} \right) \) and so an iterated integral is

\[
\int_0^{7/8} \int_x^{1 - \frac{x}{2}} \int_0^{4(1 - \frac{x}{2} - y)} dz \, dy \, dx \quad + \quad \int_0^{7/8} \int_y^{1 - \frac{x}{2}} \int_0^{4(1 - \frac{y}{2} - x)} dz \, dx \, dy = \frac{7}{6}
\]

21.6 Exercises With Answers

1. Evaluate the integral \( \int_4^7 \int_5^x \int_5^y \, dz \, dy \, dx \)

   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 \( \int_0^2 \int_0^{2-5x} \int_0^{4-4x-3y} (2x) \, dz \, dy \, dx \)

   Answer:
   \[-\frac{196}{3} \]

4. Evaluate the integral \( \int_0^8 \int_4^3 \int_4^y (x - y) \, dz \, dy \, dx \)

   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 \( \int_0^\pi \int_0^{2y} \int_0^{y+z} \sin (x + y) \, dx \, dz \, dy \)

   Answer:
   \[-\frac{19}{4} \]

7. Fill in the missing limits. \( \int_0^1 \int_0^x \int_0^z f (x, y, z) \, dx \, dy \, dz = \int_0^1 \int_0^y \int_0^z f (x, y, z) \, dx \, dy \, dz \)

   \( \int_0^1 \int_0^z \int_0^2 f (x, y, z) \, dx \, dy \, dz = \int_0^1 \int_0^1 \int_0^z f (x, y, z) \, dy \, dz \, dx \)

   \( \int_0^1 \int_0^z \int_0^z f (x, y, z) \, dx \, dy \, dz = \int_0^1 \int_0^z \int_0^z f (x, y, z) \, dy \, dz \, dx \)

   Answer:
   \( \int_0^1 \int_0^z \int_0^z f (x, y, z) \, dx \, dy \, dz = \int_0^1 \int_0^1 \int_0^z f (x, y, z) \, dy \, dz \, dx \)

   \( \int_0^1 \int_0^z \int_0^{2y} f (x, y, z) \, dx \, dy \, dz = \int_0^1 \int_0^1 \int_0^{2y} f (x, y, z) \, dy \, dz \, dx \)

   \( \int_0^1 \int_0^z \int_0^{y+z} f (x, y, z) \, dx \, dy \, dz = \int_0^1 \left[ \int_0^x \int_0^y f (x, y, z) \, dy \, dz \, dx + \int_x^1 \int_0^y f (x, y, z) \, dy \, dz \, dx \right] \, dx \)

   \( \int_0^1 \int_0^z \int_0^{y+z} f (x, y, z) \, dx \, dy \, dz = \)
Find the volume of the bounded region determined by \( \frac{1}{4}x + y + \frac{1}{4}z = 1 \) and the planes \( x = 0, y = 0, z = 0 \).
Answer: 
\[
\int_0^{1/2} \int_0^{2y} f(x, y, z) \, dx \, dz \, dy + \int_0^1 \int_0^{1/4} f(x, y, z) \, dx \, dz \, dy
\]
\[
\int_0^7 \int_2^5 \int_0^{y+z} f(x, y, z) \, dx \, dy \, dz = \int_0^3 \int_2^7 \int_0^z f(x, y, z) \, dx \, dy \, dz
\]

8. Find the volume of \( R \) where \( R \) is the bounded region formed by the plane \( \frac{1}{4}x + \frac{1}{2}y + \frac{1}{4}z = 1 \) and the planes \( x = 0, y = 0, z = 0 \).
Answer: 
\[
\int_0^5 \int_0^{1-\frac{1}{4}x} \int_0^{\frac{1}{4}x - 4y} dz \, dy \, dx = \frac{10}{3}
\]

9. Find the volume of \( R \) where \( R \) is the bounded region formed by the plane \( \frac{1}{5}x + \frac{1}{2}y + \frac{1}{4}z = 1 \) and the planes \( x = 0, y = 0, z = 0 \).
Answer: 
\[
\int_0^5 \int_0^{2-\frac{1}{4}x} \int_0^{\frac{1}{4}x - 2y} dz \, dy \, dx = \frac{20}{3}
\]

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: 
\[
\int_0^4 \int_0^{2-\frac{1}{4}x} \int_0^{\frac{1}{4}x - 2y} (y) \, dz \, dy \, dx = 2
\]

11. Find the mass of the bounded region, \( R \) formed by the plane \( \frac{1}{2}x + \frac{1}{2}y + \frac{1}{4}z = 1 \) and the planes \( x = 0, y = 0, z = 0 \) if the density is \( \rho(x, y, z) = z^2 \)
Answer: 
\[
\int_0^2 \int_0^{2-x} \int_0^{4-2x-2y} (z^2) \, dz \, dy \, dx = \frac{64}{15}
\]

12. Here is an iterated integral: \( \int_0^3 \int_0^{3-x} z^2 \, dy \, dx \). Write as an iterated integral in the following orders: \( dx \, dy \, dz, \, dx \, dz \, dy, \, dy \, dx \, dz, \, dy \, dz \, dx \).
Answer:
\[
\int_0^3 \int_0^x \int_0^{3-x} dy \, dz \, dx, \int_0^3 \int_0^{\sqrt{3}} \int_0^{3-y} dy \, dz \, dx, \int_0^3 \int_0^{3-\sqrt{2}} \int_0^{3-y} dx \, dy \, dz,
\]
\[
\int_0^3 \int_0^{\sqrt{3}} \int_0^{3-y} dz \, dx \, dy, \int_0^3 \int_0^{3-\sqrt{2}} \int_0^{3-y} dx \, dz \, dy
\]

13. Find the volume of the bounded region determined by \( 5y + 2z = 4, x = 4 - y^2, y = 0, x = 0 \).
Answer: 
\[
\int_0^{\frac{4}{2}} \int_0^{2-\frac{1}{2}y} \int_0^{4-y^2} dx \, dz \, dy = \frac{1168}{375}
\]

14. Find the volume of the bounded region determined by \( 4y + 3z = 3, x = 4 - y^2, y = 0, x = 0 \).
Answer: 
\[
\int_0^{\frac{4}{2}} \int_0^{1-\frac{1}{4}y} \int_0^{4-y^2} dx \, dz \, dy = \frac{375}{256}
\]

15. Find the volume of the bounded region determined by \( 3y + z = 3, x = 4 - y^2, y = 0, x = 0 \).
Answer: 
\[
\int_0^1 \int_0^{3-3y} \int_0^{4-y^2} dx \, dz \, dy = \frac{23}{4}
\]
16. Find the volume of the region bounded by $x^2 + y^2 = 16$, $z = 3x$, $z = 0$, and $x \geq 0$.

Answer: 
\[
\int_0^4 \int_{\sqrt{16-x^2}}^{3x} dz \, dy \, dx = 128
\]

17. Find the volume of the region bounded by $x^2 + y^2 = 25$, $z = 2x$, $z = 0$, and $x \geq 0$.

Answer: 
\[
\int_0^5 \int_{\sqrt{25-x^2}}^{2x} dz \, dy \, dx = \frac{500}{3}
\]

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: 
\[
8 \int_0^3 \int_{\sqrt{9-y^2}}^{\sqrt{9-y^2}} dz \, dx \, dy = 144
\]

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: 
\[
\int_0^4 \int_0^{\sqrt{16-x^2}} \int_0^{a^2-x^2-y^2} (z) \, dz \, dy \, dx = \frac{512}{3} \pi
\]

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: 
\[
\int_0^5 \int_0^{\sqrt{25-x^2}} \int_0^{25-x^2-y^2} (x+1) \, dz \, dy \, dx = \frac{625}{8} \pi + \frac{1250}{3}
\]

21. Find the volume of the region bounded by $x^2 + y^2 = 9$, $z = 0$, $z = 5 - y$

Answer: 
\[
\int_{-3}^{3} \int_{\sqrt{9-x^2}}^{5-y} dz \, dy \, dx = 45 \pi
\]

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: 
\[
\int_0^1 \int_0^{1-x} \int_0^{2-x-y} dz \, dy \, dx + \int_0^1 \int_0^{1-y} \int_0^{2-x-y} dz \, dx \, dy = \frac{4}{9}
\]

23. Find the volume of the bounded region determined by $x \geq 0, y \geq 0, z \geq 0$, and $\frac{1}{2}x + y + \frac{1}{3}z = 1$, and $x + \frac{1}{2}y + \frac{1}{3}z = 1$.

Answer: 
\[
\int_0^1 \int_0^{1-x} \int_0^{3-\frac{3}{2}x-3y} dz \, dy \, dx + \int_0^1 \int_0^{1-y} \int_0^{3-3x-\frac{3}{2}y} dz \, dx \, dy = \frac{7}{8}
\]

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: 
\[
\int_{-\frac{3}{5}}^{\frac{3}{5}} \int_{-\frac{1}{5}\sqrt{9-25x^2}}^{\frac{1}{5}\sqrt{9-25x^2}} \int_0^{x+2} (x) \, dz \, dy \, dx = \frac{81}{1000} \pi
\]

25. Find 
\[
\int_0^1 \int_0^{35-5z} \int_{\frac{1}{2}x}^{7-z} (7 - z) \, dy \, dz \, dx.
\]

Answer: 
You need to interchange the order of integration. 
\[
\int_0^1 \int_0^{7-z} \int_0^{5y} (7 - z) \cos (y^2) \, dy \, dz = \frac{2}{9} \cos 36 - \frac{2}{9} \cos 49
\]
26. Find \( \int_0^2 \int_0^{12-3z} \int_0^4 z (4 - z) \exp (y^2) \, dy \, dx \, dz. \)

Answer:
You need to interchange the order of integration. 
\( \int_0^2 \int_0^4 \int_0^{12-3z} z (4 - z) \exp (y^2) \, dy \, dx \, dz = -\frac{3}{4} e^4 - 9 + \frac{3}{4} e^{16} \)

27. Find \( \int_0^2 \int_0^{25-5z} \int_0^5 z (5 - z) \exp (x^2) \, dx \, dy \, dz. \)

Answer:
You need to interchange the order of integration.
\( \int_0^2 \int_0^{5-z} \int_0^{5x} (5 - z) \exp (x^2) \, dy \, dx \, dz = -\frac{5}{4} e^9 - 20 + \frac{5}{4} e^{25} \)

28. Find \( \int_0^1 \int_0^{10-2z} \int_0^{5-z} \sin x \, dx \, dy \, dz. \)

Answer:
You need to interchange the order of integration.
\( \int_0^1 \int_0^{5-z} \int_0^{2x} \sin x \, dy \, dx \, dz = -2 \sin 1 \cos 5 + 2 \cos 1 \sin 5 + 2 - 2 \sin 5 \)

29. Find \( \int_0^{20} \int_0^2 \int_0^{6-z} \sin x \, dx \, dy \, dz + \int_0^{30} \int_0^{6-z} \int_0^1 \sin x \, dx \, dy \, dz. \)

Answer:
You need to interchange the order of integration.
\( \int_0^2 \int_0^{30-5z} \int_0^4 \sin x \, dx \, dy \, dz = \int_0^2 \int_0^{6-z} \int_0^{5x} \sin x \, dy \, dx \, dz = -5 \sin 2 \cos 6 + 5 \cos 2 \sin 6 + 10 - 5 \sin 6 \)
Chapter 22

The Integral In Other Coordinates

22.1 Cylindrical And Spherical Coordinates

22.1.1 Geometric Description

Cylindrical coordinates are defined as follows.

\[ \mathbf{x}(r, \theta, z) \equiv \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r \cos(\theta) \\ r \sin(\theta) \\ z \end{pmatrix}, \]

\[ r \geq 0, \theta \in [0, 2\pi), z \in \mathbb{R} \]

Spherical coordinates are a little harder. Recall these are given by

\[ \mathbf{x}(\rho, \theta, \phi) \equiv \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \rho \sin(\phi) \cos(\theta) \\ \rho \sin(\phi) \sin(\theta) \\ \rho \cos(\phi) \end{pmatrix}, \]

\[ \rho \geq 0, \theta \in [0, 2\pi), \phi \in [0, \pi] \]

The following picture relates the various coordinates.

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 labelled as \((x_1, y_1, z_1), (r, \theta, z_1)\),
CHAPTER 22. THE INTEGRAL IN OTHER COORDINATES

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 \(r = \rho \sin(\phi)\) and is the usual polar coordinate while \(\theta\) is the other polar coordinate. 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 of “latitude” and finally, \(\theta\) determines which point is on this circle by specifying a circle of “longitude”. 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. Note that \(\theta\) is the same in the two coordinate systems and that \(\rho \sin \phi = r\).

### 22.1.2 Volume And Integrals In Cylindrical Coordinates

The increment of three dimensional volume in cylindrical coordinates is \(dV = r dr d\theta dz\). It is just a chunk of two dimensional area, \(r dr d\theta\) times the height \(dz\) which gives three dimensional volume. Here is an example.

**Example 22.1.1** Find the volume of the three dimensional region between the graphs of \(z = 4 - 2y^2\) and \(z = 4x^2 + 2y^2\). Next, if \(R\) is this region, find \(\int_R z dV\).

Where do the two surfaces intersect? This happens when \(4x^2 + 2y^2 = 4 - 2y^2\) which is the curve in the \(xy\) plane, \(x^2 + y^2 = 1\). Thus \((x, y)\) is on the inside of this circle while \(z\) goes from \(4x^2 + 2y^2\) to \(4 - 2y^2\). Denoting the unit disk by \(D\), the desired integral is

\[
\int_D \int_{4x^2+2y^2}^{4-2y^2} dz dA
\]

I will use the \(dA\) which corresponds to polar coordinates so this will then be in cylindrical coordinates. Thus the above equals

\[
\int_0^{2\pi} \int_0^1 r^2 \sin^2(\theta) d\theta dA = 2\pi
\]

Note this is really not much different than simply using polar coordinates to integrate the difference of the two values of \(z\) This is

\[
\int_D (4 - 2y^2 - (4x^2 + 2y^2)) dA = \int_D (4 - 4r^2) dA = \int_0^{2\pi} \int_0^1 (4 - 4r^2) r dr d\theta = 2\pi
\]
Now consider the integral. Here we can’t conveniently reduce to polar coordinates but we can do the integral as follows.

\[
\int_R z \, dV = \int_0^{2\pi} \int_0^1 \int_0^{4\cos^2(\theta) + 2\sin^2(\theta)} z \, dz \, r \, dr \, d\theta
\]

This will be a long and tedious computation but it is perfectly well defined. Here is another example.

**Example 22.1.2** Find the volume of the three dimensional region between the graphs of \( z = 0, z = \sqrt{x^2 + y^2}, \) and the cylinder \((x - 1)^2 + y^2 = 1).\)

Consider the cylinder. It reduces to \( r^2 = 2r \cos \theta \) or more simply \( r = 2 \cos \theta \). This is the graph of a circle having radius 1 and centered at \((1,0)\). Therefore, \( \theta \in [-\pi/2, \pi/2] \). It follows that the cylindrical coordinate description of this volume is

\[
\int_{-\pi/2}^{\pi/2} \int_0^{2 \cos \theta} \int_0^r dz \, r \, dr \, d\theta = \frac{32}{9}
\]

You need to always worry about the limits of the integrals when you do these problems.

### 22.1.3 Volume And Integrals In Spherical Coordinates

What is the increment of volume in spherical coordinates? There are two ways to see what this is, through art and through a systematic procedure. First consider art. Here is a picture.

In the picture there are two concentric spheres formed by making \( \rho \) two different constants and surfaces which correspond to \( \theta \) assuming two different constants and \( \phi \) assuming two different constants. These intersecting surfaces form the little box in the picture. Here is a more detailed blow up of the little box.
What is the volume of this little box? Length \( \approx \rho d\phi \), width \( \approx \rho \sin(\phi) d\theta \), height \( \approx d\rho \) and so the volume increment for spherical coordinates is

\[
dV = \rho^2 \sin(\phi) d\rho d\theta d\phi
\]

Now what is really going on? Consider the dot in the picture of the little box. Fixing \( \theta \) and \( \phi \) at their values at this point and differentiating with respect to \( \rho \) leads to a little vector of the form

\[
\begin{pmatrix}
\sin(\phi) \\
\sin(\phi) \\
\cos(\phi)
\end{pmatrix}
\]

which points out from the surface of the sphere. Next keeping \( \rho \) and \( \theta \) constant and differentiating only with respect to \( \phi \) leads to an infinitesimal vector in the direction of a line of longitude,

\[
\begin{pmatrix}
\rho \cos(\phi) \\
\rho \sin(\phi) \\
-\rho \sin(\phi)
\end{pmatrix}
\]

and finally keeping \( \rho \) and \( \phi \) constant and differentiating with respect to \( \theta \) leads to the third infinitesimal vector which points in the direction of a line of latitude.

\[
\begin{pmatrix}
-\rho \sin(\phi) \\
\rho \sin(\phi) \\
0
\end{pmatrix}
\]

To find the increment of volume, we just need to take the absolute value of the determinant which has these vectors as columns, (Remember this is the absolute value of the box product.) exactly as was the case for polar coordinates. This will also yield

\[
dV = \rho^2 \sin(\phi) d\rho d\theta d\phi.
\]

However, in contrast to the drawing of pictures, this procedure is completely general and will handle all curvilinear coordinate systems and in any dimension. This is discussed more later.
Example 22.1.3 Find the volume of a ball, $B_R$ of radius $R$. Then find $\int_{B_R} z^2 dV$ where $z$ is the rectangular $z$ coordinate of a point.

In this case, $U = (0, R] \times [0, \pi] \times [0, 2\pi)$ and use spherical coordinates. Then this 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

$$\int_{B_R} 1 dV = \int_U \rho^2 \sin \phi \, dV$$

$$= \int_0^R \int_0^\pi \int_0^{2\pi} \rho^2 \sin \phi \, d\theta \, d\phi \, d\rho = \frac{4}{3} R^3 \pi.$$  

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.

As for the integral, it is no harder to set up. You know from the transformation equations that $z = \rho \cos \phi$. Then you want

$$\int_{B_R} z dV = \int_0^R \int_0^\pi \int_0^{2\pi} (\rho \cos \phi)^2 \rho^2 \sin \phi \, d\theta \, d\phi \, d\rho = \frac{4}{15} \pi R^5.$$  

This will be pretty easy also although somewhat more messy because the function you are integrating is not just 1 as it is when you find the volume.

Example 22.1.4 A cone is cut out of a ball of radius $R$ as shown in the following picture, the diagram on the left being a side view. The angle of the cone is $\pi/3$. Find the volume of what is left.

Use spherical coordinates. This volume is then

$$\int_{\pi/6}^{\pi} \int_0^{2\pi} \int_0^R \rho^2 \sin(\phi) \, d\rho \, d\theta \, d\phi = \frac{2}{3} \pi R^3 + \frac{1}{3} \sqrt{3} \pi R^3.$$  

Now change the example a little by cutting out a cone at the bottom which has an angle of $\pi/2$ as shown. What is the volume of what is left?
This time you would have the volume equals
\[ \int_{\pi/6}^{3\pi/4} \int_0^{2\pi} \int_0^R \rho^2 \sin(\phi) \, d\rho \, d\theta \, d\phi = \frac{1}{3} \sqrt{2\pi} R^3 + \frac{1}{3} \sqrt{3\pi} R^3 \]

**Example 22.1.5** Next suppose the ball of radius \( R \) is a sort of an orange and you remove a slice as shown in the picture. What is the volume of what is left? Assume the slice is formed by the two half planes \( \theta = 0 \) and \( \theta = \pi/4 \).

Using spherical coordinates, this gives for the volume
\[ \int_{\pi/4}^\pi \int_0^{2\pi} \int_0^R \rho^2 \sin(\phi) \, d\rho \, d\theta \, d\phi = \frac{7}{6} \pi R^3 \]

**Example 22.1.6** Now remove the same two cones as in the above examples along with the same slice and find the volume of what is left. Next, if \( R \) is the region just described, find \( \int_R x \, dV \).

This time you need
\[ \int_{\pi/6}^{3\pi/4} \int_{\pi/4}^{2\pi} \int_0^R \rho^2 \sin(\phi) \, d\rho \, d\theta \, d\phi = \frac{7}{24} \sqrt{2\pi} R^3 + \frac{7}{24} \sqrt{3\pi} R^3 \]

As to the integral, it equals
\[ \int_{\pi/6}^{3\pi/4} \int_{\pi/4}^{2\pi} \int_0^R (\rho \sin(\phi) \cos(\theta)) \rho^2 \sin(\phi) \, d\rho \, d\theta \, d\phi = -\frac{1}{192} \sqrt{2} R^4 \left(7\pi + 3\sqrt{3} + 6\right) \]

This is because, in terms of spherical coordinates, \( x = \rho \sin(\phi) \cos(\theta) \).

**Example 22.1.7** Set up the integrals to find the volume of the cone \( 0 \leq z \leq 4, z = \sqrt{x^2 + y^2} \). Next, if \( R \) is the region just described, find \( \int_R z \, dV \).
This is entirely the wrong coordinate system to use for this problem but it is a good exercise. Here is a side view.

You need to figure out what $\rho$ is as a function of $\phi$ which goes from 0 to $\pi/4$. You should get

$$\int_0^{2\pi} \int_0^{\pi/4} \int_0^{4 \sec(\phi)} \rho^2 \sin(\phi) \, d\rho d\phi d\theta = \frac{64}{3} \pi$$

As to $\int_R z \, dV$, it equals

$$\int_0^{2\pi} \int_0^{\pi/4} \int_0^{4 \sec(\phi)} z \rho \cos(\phi) \rho \sin(\phi) \, d\rho d\phi d\theta = 64\pi$$

Example 22.1.8 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

$$\det \begin{pmatrix} \cos \theta & -r \sin \theta & 0 \\ \sin \theta & r \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix} = r.$$ 

It follows the volume element in cylindrical coordinates is $r \, d\theta \, dr \, dz$.

Example 22.1.9 In the cone of Example 22.1.7 set up the integrals for finding the volume in cylindrical coordinates.

This is a better coordinate system for this example than spherical coordinates. This time you should get

$$\int_0^{2\pi} \int_0^{4} \int_0^{r} rdzdrd\theta = \frac{64}{3} \pi$$

Example 22.1.10 This example uses spherical coordinates to verify an important conclusion about gravitational force. Let the hollow sphere, $H$ be defined by $a^2 < x^2 + y^2 + z^2 < b^2$
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 that the force of gravity acting on this unit mass is \(\left(\alpha G \int_H \frac{(z-z_0)}{\left\{x^2 + y^2 + (z-z_0)^2\right\}^{3/2}} \ dV\right) \mathbf{k}\) and then show that if \(|z_0| > b\) 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 \(|z_0| < a\), the total force acting on the point mass from gravity equals zero. Here \(G\) is the gravitation constant and \(\alpha\) 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.\(^1\)

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

\[
\frac{x\mathbf{i} + y\mathbf{j} + (z-z_0)\mathbf{k}}{|x\mathbf{i} + y\mathbf{j} + (z-z_0)\mathbf{k}|} \frac{1}{\left(x^2 + y^2 + (z-z_0)^2\right)} \alpha G \ dV = \\
(x\mathbf{i} + y\mathbf{j} + (z-z_0)\mathbf{k}) \frac{1}{\left(x^2 + y^2 + (z-z_0)^2\right)^{3/2}} \alpha G \ dV
\]

Therefore, the total force is

\[
\int_H (x\mathbf{i} + y\mathbf{j} + (z-z_0)\mathbf{k}) \frac{1}{\left(x^2 + y^2 + (z-z_0)^2\right)^{3/2}} \alpha G \ dV.
\]

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

\[
\alpha G \mathbf{k} \int_H \frac{(z-z_0)}{\left(x^2 + y^2 + (z-z_0)^2\right)^{3/2}} \ dV
\]

as claimed. Now for the interesting part, the integral is evaluated. In spherical coordinates this integral is.

\[
\int_0^{2\pi} \int_0^a \int_0^\pi \frac{(\rho \cos \phi - z_0)}{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)^{3/2}} \rho^2 \sin \phi \ d\phi \ d\rho \ d\theta. \tag{22.1}
\]

Rewrite the inside integral and use integration by parts to obtain this inside integral equals

\[
\frac{1}{2z_0} \int_0^\pi \left(\rho^2 \cos \phi - \rho z_0\right) \frac{(2z_0 \rho \sin \phi)}{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)^{3/2}} \ d\phi = \\
\frac{1}{2z_0} \left(-\frac{-\rho^2 - \rho z_0}{\sqrt{\rho^2 + z_0^2 + 2\rho z_0}} + 2 \frac{\rho^2 - \rho z_0}{\sqrt{\rho^2 + z_0^2 - 2\rho z_0}} - \int_0^\pi \frac{2\rho^2 \sin \phi}{\sqrt{\rho^2 + z_0^2 - 2\rho z_0 \cos \phi}} \ d\phi\right). \tag{22.2}
\]

There are some cases to consider here.

\(^1\)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
\[
\frac{2\rho (\rho + z_0)}{(\rho + z_0)^2} + \frac{2\rho (\rho - z_0)}{(\rho - z_0)^2} = \frac{2\rho (\rho + z_0)}{(\rho + z_0)} + \frac{2\rho (\rho - z_0)}{\rho - z_0} = 4\rho
\]
and so the expression in \( 22.2 \) equals
\[
\frac{1}{2z_0^2} \left( 4\rho - \int_0^\pi 2\rho^2 \frac{\sin \phi}{\sqrt{\rho^2 + z_0^2 - 2\rho z_0 \cos \phi}} d\phi \right)
\]
\[
= \frac{1}{2z_0} \left( 4\rho - \frac{1}{z_0} \int_0^\pi \rho \frac{2\rho z_0 \sin \phi}{\sqrt{\rho^2 + z_0^2 - 2\rho z_0 \cos \phi}} d\phi \right)
\]
\[
= \frac{1}{2z_0} \left( 4\rho - \frac{2\rho}{z_0} (\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)^{1/2} \right|_0^\pi
\]
\[
= \frac{1}{2z_0} \left( 4\rho - \frac{2\rho}{z_0} [\rho + z_0 - (\rho - z_0)] \right) = 0.
\]
Therefore, in this case the inner integral of \( 22.1 \) 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 \( 22.2 \) are
\[
\frac{2\rho (\rho + z_0)}{(\rho + z_0)^2} + \frac{2\rho (\rho - z_0)}{(\rho - z_0)^2} = \frac{2\rho (\rho + z_0)}{(\rho + z_0)} + \frac{2\rho (\rho - z_0)}{\rho - z_0} = 0.
\]
Therefore in this case, \( 22.2 \) equals
\[
\frac{1}{2z_0^2} \left( - \int_0^\pi 2\rho^2 \frac{\sin \phi}{\sqrt{\rho^2 + z_0^2 - 2\rho z_0 \cos \phi}} d\phi \right)
\]
\[
= -\frac{\rho}{2z_0^2} \left( \int_0^\pi \frac{2\rho z_0 \sin \phi}{\sqrt{\rho^2 + z_0^2 - 2\rho z_0 \cos \phi}} d\phi \right)
\]
which equals
\[
\frac{-\rho}{z_0^2} \left( (\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)^{1/2} \right|_0^\pi
\]
\[
= \frac{-\rho}{z_0^2} [\rho + z_0 - (z_0 - \rho)] = \frac{-2\rho^2}{z_0^2}.
\]
Thus the inner integral of \( 22.1 \) reduces to the above simple expression. Therefore, \( 22.1 \) equals
\[
\int_0^{2\pi} \int_a^b \left( -\frac{2}{z_0^2} \rho^2 \right) d\rho d\theta = \frac{4}{3} \pi \frac{b^3 - a^3}{z_0^2}
\]
and so
\[
\alpha G k \int_H \left[ \frac{z - z_0}{x^2 + y^2 + (z - z_0)^2} \right]^{3/2} dV
\]
\[
= \alpha G k \left( -\frac{4}{3} \pi \frac{b^3 - a^3}{z_0^2} \right) = -kG \frac{\text{total mass}}{z_0^2}.
\]
22.2 Exercises

1. Find the volume of the region bounded by \( z = 0, x^2 + (y - 2)^2 = 4 \), and \( z = \sqrt{x^2 + y^2} \).

2. Find the volume of the region, \( z \geq 0, x^2 + y^2 \leq 4 \), and \( z \leq 4 - \sqrt{x^2 + y^2} \).

3. Find the volume of the region which is between the surfaces \( z = 5y^2 + 9x^2 \) and \( z = 9 - 4y^2 \).

4. Find the volume of the region which is between \( z = x^2 + y^2 \) and \( z = 5 - 4x \). \textbf{Hint:} You might want to change variables at some point.

5. The ice cream in a sugar cone is described in spherical coordinates by \( \rho \in [0, 10], \phi \in [0, \frac{1}{3} \pi], \theta \in [0, 2\pi] \). If the units are in centimeters, find the total volume in cubic centimeters of this ice cream.

6. Find the volume between \( z = 3 - x^2 - y^2 \) and \( z = 2\sqrt{x^2 + y^2} \).

7. 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?

8. Find the volume of the cone defined by \( z \in [0, 4] \) having angle \( \pi/2 \). Use spherical coordinates.

9. 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 \( z = 2 \). Write integrals for the mass of what is left. In spherical coordinates and in cylindrical coordinates.

10. A ball of radius 4 has a cone taken out of the top which has an angle of \( \pi/2 \) and then a cone taken out of the bottom which has an angle of \( \pi/3 \). Then a slice, \( \theta \in [0, \pi/4] \) is removed. What is the volume of what is left?

11. In Example \( \text{Example 22.1.10} \) on Page 443 check out all the details by working the integrals to be sure the steps are right.

12. What if the hollow sphere in Example \( \text{Example 22.1.10} \) were in two dimensions and everything, including Newton’s law still held? Would similar conclusions hold? Explain.
13. Convert the following integrals into integrals involving cylindrical coordinates and then evaluate them.

(a) \( \int_{-2}^{2} \int_{0}^{\sqrt{4-x^2}} \int_{0}^{x} ydzdydx \)

(b) \( \int_{-1}^{1} \int_{0}^{\sqrt{1+y^2}} \int_{0}^{x+y} dzdxy \)

(c) \( \int_{0}^{1} \int_{0}^{\sqrt{1-x^2}} \int_{z}^{1} dzdxy \)

(d) For \( a > 0 \), \( \int_{-a}^{a} \int_{-\sqrt{a^2-x^2}}^{\sqrt{a^2-x^2}} \int_{-\sqrt{a^2-x^2-y^2}}^{\sqrt{a^2-x^2-y^2}} dzdxy \)

(e) \( \int_{-1}^{1} \int_{\sqrt{1-x^2}}^{\sqrt{4-x^2-y^2}} dzdxy \)

14. Convert the following integrals into integrals involving spherical coordinates and then evaluate them.

(a) \( \int_{-\alpha}^{\alpha} \int_{-\sqrt{\alpha^2-x^2}}^{\sqrt{\alpha^2-x^2}} \int_{-\sqrt{\alpha^2-x^2-y^2}}^{\sqrt{\alpha^2-x^2-y^2}} dzdxy \)

(b) \( \int_{-1}^{1} \int_{0}^{\sqrt{1-x^2}} \int_{0}^{\sqrt{1-x^2-y^2}} dzdxy \)

(c) \( \int_{-\sqrt{2}}^{\sqrt{2}} \int_{-\sqrt{2-x^2}}^{\sqrt{2-x^2}} \int_{-\sqrt{x^2+y^2}}^{\sqrt{x^2+y^2}} dzdxy \)

(d) \( \int_{-\sqrt{3}}^{\sqrt{3}} \int_{-\sqrt{3-x^2}}^{\sqrt{3-x^2}} \int_{0}^{\sqrt{4-x^2-y^2}} dzdxy \)

(e) \( \int_{-1}^{1} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \int_{-\sqrt{4-x^2-y^2}}^{\sqrt{4-x^2-y^2}} dzdxy \)

22.3 The General Procedure

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 \( dx_1, \ldots, dx_n \) so \( dV = dx_1 \, dx_2 \cdots dx_n \) 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. Suppose for example that \( x_i = f_i(u_1, u_2, u_3) \) for \( i = 1, 2, 3 \). For example, you could have cylindrical or spherical coordinates. We need a description of a little chunk of volume in terms of the \( u_i \). Consider the following picture.
The sides of the parallelepiped are \( \frac{\partial \mathbf{x}}{\partial u_i} \, du_i \) for \( i = 1, 2, 3 \). Also as the \( du_i \) get very small, the curvy parallelepiped and the straight one would seem to be close in volume. Therefore, the increment of volume in terms of the coordinates \( u_1, u_2, u_3 \) is just the absolute value of the box product of these straight vectors,

\[
\left[ \frac{\partial \mathbf{x}}{\partial u_1} \right] du_1, \left[ \frac{\partial \mathbf{x}}{\partial u_2} \right] du_2, \left[ \frac{\partial \mathbf{x}}{\partial u_3} \right] du_3.
\]

Recall that this is just

\[
\left| \text{det} \left( \frac{\partial \mathbf{x}}{\partial u_1}, \frac{\partial \mathbf{x}}{\partial u_2}, \frac{\partial \mathbf{x}}{\partial u_3} \right) \right| du_1 du_2 du_3.
\]

The situation is exactly the same in two dimensions. There you would have the increment of area would be the following norm. Recall the geometric definition of the cross product.

\[
\left| \begin{pmatrix} x_{u_1}(u_1, u_2) \\ y_{u_1}(u_1, u_2) \\ 0 \end{pmatrix} \times \begin{pmatrix} x_{u_2}(u_1, u_2) \\ y_{u_2}(u_1, u_2) \\ 0 \end{pmatrix} \right| du_1 du_2
\]

where in this special case, we denote \( \mathbf{x}(u_1, u_2) \equiv \begin{pmatrix} x(u_1, u_2) \\ y(u_1, u_2) \end{pmatrix} \). So what is that cross product? It equals

\[
\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x_{u_1}(u_1, u_2) & y_{u_1}(u_1, u_2) & 0 \\ x_{u_2}(u_1, u_2) & y_{u_2}(u_1, u_2) & 0 \end{vmatrix}
\]

and this equals

\[
(x_{u_1}y_{u_2} - x_{u_2}y_{u_1}) \mathbf{k}
\]

Thus the norm of this is just the norm of the above vector which equals \( |(x_{u_1}y_{u_2} - x_{u_2}y_{u_1})| = |\text{det} \left( \frac{\partial \mathbf{x}}{\partial u_1}, \frac{\partial \mathbf{x}}{\partial u_2} \right) | du_1 du_2 \)

which is exactly the same pattern. In fact, this pattern goes on working for any dimension and gives a systematic way to compute the appropriate increment of \( n \) dimensional volume for any \( n \).
Definition 22.3.1 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

\[
\det \left( \frac{\partial x(u)}{\partial u_1}, \cdots, \frac{\partial x(u)}{\partial u_n} \right) = \frac{\partial (x_1, \cdots, x_n)}{\partial (u_1, \cdots, u_n)}
\]

Also, the symbol,

\[
\left| \frac{\partial (x_1, \cdots, x_n)}{\partial (u_1, \cdots, u_n)} \right| \, du_1 \cdots du_n
\]

called the volume element or increment of volume or increment of area.

\[
\det \left( \frac{\partial x(u)}{\partial u_1}, \cdots, \frac{\partial x(u)}{\partial u_n} \right) = \frac{\partial (x_1, \cdots, x_n)}{\partial (u_1, \cdots, u_n)}
\]

is called the Jacobian.

This has given motivation for the following fundamental procedure often called the change of variables formula which holds under fairly general conditions.

Procedure 22.3.2 Suppose \( U \) is an open subset of \( \mathbb{R}^n \) for \( n > 0 \) and suppose \( f : U \to f(U) \) is a \( C^1 \) function which is one to one, \( x = f(u) \). Then if \( h : f(U) \to \mathbb{R} \),

\[
\int_U h(f(u)) \left| \frac{\partial (x_1, \cdots, x_n)}{\partial (u_1, \cdots, u_n)} \right| \, dV = \int_{f(U)} h(x) \, dV.
\]

Use this procedure to compute the volume element in spherical coordinates. This is left as an exercise.

Example 22.3.3 Find the Jacobians of the following transformations.

1. \( x = 5u - v, y = u + 3v \)
   
   \[
   (a) \begin{vmatrix} 5 & -1 \\ 1 & 3 \end{vmatrix} = 16
   \]

2. \( x = uv, y = u/v \)
   
   \[
   (a) \begin{vmatrix} v & u \\ 1/v & -u/v^2 \end{vmatrix} = -2\frac{u}{v}
   \]

3. \( x = v + w^2, y = w + u^2, z = u + v^2 \)
   
   \[
   (a) \begin{vmatrix} 0 & 1 & 2w \\ 2u & 0 & 1 \\ 1 & 2v & 0 \end{vmatrix} = 8uvw + 1
   \]

\[\text{This will cause non overlapping infinitesimal boxes in } U \text{ to be mapped to non overlapping infinitesimal parallelepipeds in } V.\]

Also, in the context of the Riemann integral we should say more about the set \( U \) 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 in this part of the book, I think it best to give a memorable version of it which is essentially correct in all examples of interest.
Example 22.3.4 Let \( S = \{(u, v) : 0 \leq u \leq 3, 0 \leq v \leq 2\} \). Thus \( S \) is a rectangle. Let \( x = 2u + 3v, y = u - v \). Describe the image of \( S \) under this transformation. **Hint:** It will consist of some sort of quadrilateral. Thus you will know what it is if you give its vertices.

The vertices of \( S \) are \((0, 0), (3, 0), (3, 2), (0, 2) = (u, v) \). Then these become \((0, 0), (6, 3), (12, 1), (6, -2) \).

Example 22.3.5 Find the area of the bounded region \( R \), determined by \( 5x + y = 1, 5x + y = 9, y = 2x, \) and \( y = 5x \).

Here you let \( u = 5x + y, v = y/x \). Then you have \((u, v) \in [1, 9) \times [2, 5] \). Now you need to solve for \( x, y \) in terms of \( u, v \).

\[ u = 5x + y, \text{ Thus } x = \frac{u}{v + 5} \]

Then the Jacobian

\[ \left| \begin{array}{cc} \frac{1}{v + 5} & -\frac{u}{(v + 5)^2} \\ \\ \frac{v}{v + 5} & \frac{5u}{(v + 5)^2} \end{array} \right| = \frac{u}{(v + 5)^2} \]

The area is just

\[ \int_2^5 \int_1^9 \frac{u}{(v + 5)^2} dudv = \frac{12}{7} \]

Example 22.3.6 A solid, \( R \) is determined by \( 3x + y = 2, 3x + y = 4, y = x, \) and \( y = 2x \) and the density is \( \rho = x \). Find the total mass of \( R \). Then determine \( \bar{x} \) the \( x \) coordinate of the center of mass.

You let \( u = 3x + y \) and \( v = y/x \). Thus

\[ x = \frac{u}{v + 3}, y = u \frac{x}{v + 3} \]

and \((u, v) \in [2, 4] \times [1, 2] \). Now you find the Jacobian.

\[ \left| \begin{array}{cc} \frac{1}{v + 3} & -\frac{u}{(v + 3)^2} \\ \\ \frac{v}{v + 3} & \frac{3u}{(v + 3)^2} \end{array} \right| = \frac{u}{(v + 3)^2} \]

Then the integral you want for the mass is just

\[ \int_1^2 \int_2^4 \frac{u}{v + 3} \frac{u}{(v + 3)^2} dudv = \frac{21}{100} \]

\[ \bar{x} = \frac{\int_1^2 \int_2^4 \left( \frac{u}{v + 3} \right)^2 \frac{u}{(v + 3)^2} dudv}{21/100} = \frac{61}{84} \]
Example 22.3.7 Find the volume of the region \( E \), bounded by the ellipsoid, \( \frac{1}{4}x^2 + y^2 + z^2 = 1 \).

You ought to let \( u = \frac{x}{2}, v = y, z = w \). Then \((u, v, w)\) is in the ball of radius 1 called \( B \). The Jacobian is obviously 2 and so we have

\[
\int_E dV = \int_B 2dV = 2 \left( \frac{4}{3} \pi \right) = \frac{8}{3} \pi.
\]

Example 22.3.8 Here are three vectors. \((4, 1, 2)^T, (5, 0, 2)^T, \) and \((3, 1, 3)^T\). These vectors determine a parallelepiped, \( R \), which is occupied by a solid having density \( \rho = x \). Find the mass of this solid.

You have

\[
\begin{pmatrix}
  x \\
  y \\
  z
\end{pmatrix} = u \begin{pmatrix} 4 \\ 1 \\ 2 \end{pmatrix} + v \begin{pmatrix} 5 \\ 0 \\ 2 \end{pmatrix} + w \begin{pmatrix} 3 \\ 1 \\ 3 \end{pmatrix}
\]

where \((u, v, w)\) is in \([0, 1] \times [0, 1] \times [0, 1] \). Also the Jacobian is clearly

\[
\det \begin{pmatrix} 4 & 5 & 3 \\ 1 & 0 & 1 \\ 2 & 2 & 3 \end{pmatrix} = -7
\]

and so the area increment is \( 7dudvdw \) and therefore, the mass of the solid is

\[
\int_0^1 \int_0^1 \int_0^1 x dudvdw = \int_0^1 \int_0^1 \int_0^1 (4u + 5v + 3w) 7dudvdw = 42
\]

Example 22.3.9 Find \( \int_R x^2 dV \) where \( R \) is the region bounded by the ellipsoid \( \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \). Your answer will depend on \( a, b, c \) each of which is a positive number.

You let \( u = \frac{x}{a}, v = \frac{y}{b}, w = \frac{z}{c} \) and then \((u, v, w)\) is in the ball of radius 1 \( B \). Then also you have the Jacobian equals \( abc \) and so you have

\[
\int_E x^2 dV = \int_B (au)^2 abcdV = abc \int_0^\pi \int_0^{2\pi} \int_0^1 (a^2 \rho^2 \sin(\phi))^2 \rho^2 \sin(\phi) d\rho d\theta d\phi
\]

\[
= \frac{8}{15} \pi a^3 bc
\]

Example 22.3.10 Find \( \int_R xy dA \) where \( R \) is the region in the first quadrant bounded by the lines \( y = x, y = 2x, \) and the hyperbolas \( xy = 1, xy = 2 \).
Try the transformations determined by \( u = xy \) and \( v = y/x \). Thus \((u, v) \in [1, 2] \times [1, 2]\).

First solve for \( x, y \) in terms of \( u, v \).

\[
x = \frac{1}{v} \sqrt{uv}, \quad y = \sqrt{uv}
\]

Then the Jacobian is

\[
\det \left( \begin{array}{cc}
\frac{1}{2} \sqrt{uv} & -\frac{1}{2} \frac{u}{v} \\
\frac{1}{2} \frac{v}{\sqrt{uv}} & \frac{1}{2} \frac{u}{v}
\end{array} \right) = \frac{1}{2v}
\]

Then you have

\[
\int_R xy\,dA = \int_1^2 \int_1^2 \frac{1}{v} \sqrt{uv} \sqrt{uv} \frac{1}{2v} \,dudv = \int_1^2 \int_1^2 \frac{1}{2v} \,dudv
\]

\[
= \frac{3}{4} \ln 2
\]

**Example 22.3.11** Find \( \int_R \cos \left( \frac{x - y}{x + y} \right) \,dA \) where \( R \) is bounded by straight segments joining the points \((1, 0), (2, 0), (0, 2), (0, 1)\).

Try something like \( u = x + y, \, v = x \). Then to find the new region in the \( uv \) plane, you might look at the vertices. \((x, y) \rightarrow (u, v)\) as follows.

\((1, 0) \rightarrow (1, 1), (2, 0) \rightarrow (2, 2), (0, 2) \rightarrow (2, 0), (0, 1) \rightarrow (1, 0)\)

Then you have

\[
\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}
\]

and so

\[
\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} v \\ u - v \end{pmatrix}
\]

Then the Jacobian is just

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

and so the area increment is just \( dudv \). So where do \( u, v \) live? Look at the vertices for \((u, v)\) they are \((1, 0), (2, 0), (2, 2)\), and \((1, 1)\). Thus you would have \( u \in [1, 2], 0 \leq v \leq u \).

\[
\int_R \cos \left( \frac{x - y}{x + y} \right) \,dA = \int_1^2 \int_0^u \cos \left( \frac{v - (u - v)}{u} \right) \,dvdu
\]

\[
= \int_1^2 \int_0^u \cos \left( \frac{1}{u} (2v - u) \right) \,dvdu
\]

\[
= \frac{3}{2} \sin 1
\]

### 22.4 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 \( u = \frac{y}{x}, v = 3x + 3y \). Then solving these equations for \( x \) and \( y \) yields
\[
\begin{align*}
  x &= \frac{1}{3} \frac{v}{1 + u}, \\
  y &= \frac{1}{3} \frac{u}{1 + u} 
\end{align*}
\]

Now
\[
\frac{\partial (x, y)}{\partial (u, v)} = \det \begin{pmatrix}
  -\frac{1}{3} \frac{v}{(1 + u)^2} & \frac{1}{3} \frac{3 + 3u}{(1 + u)^2} \\
  \frac{1}{3} \frac{v}{(1 + u)^2} & \frac{1}{3} \frac{3 + 3u}{(1 + u)}
\end{pmatrix} = -\frac{1}{9} \frac{v}{(1 + u)^2}.
\]

Also, \( u \in [3, 4] \) while \( v \in [1, 8] \). Therefore,
\[
\int_R dV = \int_3^4 \int_1^8 \left| -\frac{1}{9} \frac{v}{(1 + u)^2} \right| dv du = \frac{7}{40}.
\]

2. Find the area of the bounded region, \( R \), determined by \( 5x + y = 1, 5x + y = 9, y = 2x, \) and \( y = 5x \).

Answer:
Let \( u = \frac{y}{x}, v = 5x + y \). Then solving these equations for \( x \) and \( y \) yields
\[
\begin{align*}
  x &= \frac{v}{5 + u}, \\
  y &= \frac{u}{5 + u} 
\end{align*}
\]

Now
\[
\frac{\partial (x, y)}{\partial (u, v)} = \det \begin{pmatrix}
  -\frac{v}{(5 + u)^2} & \frac{1}{5 + u} \\
  \frac{u}{(5 + u)^2} & \frac{v}{5 + u}
\end{pmatrix} = -\frac{v}{(5 + u)^2}.
\]

Also, \( u \in [2, 5] \) while \( v \in [1, 9] \). Therefore,
\[
\int_R dV = \int_2^5 \int_1^9 \left| -\frac{v}{(5 + u)^2} \right| dv du = \frac{12}{7}.
\]

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 \( u = \frac{y}{x}, v = 5x + 3y \). Then solving these equations for \( x \) and \( y \) yields
\[
\begin{align*}
  x &= \frac{v}{5 + 3u}, \\
  y &= \frac{u}{5 + 3u} 
\end{align*}
\]

Now
\[
\frac{\partial (x, y)}{\partial (u, v)} = \det \begin{pmatrix}
  -\frac{3v}{(5 + 3u)^2} & \frac{1}{5 + 3u} \\
  \frac{v}{(5 + 3u)^2} & \frac{u}{5 + 3u}
\end{pmatrix} = -\frac{v}{(5 + 3u)^2}.
\]

Also, \( u \in [2, 5] \) while \( v \in [4, 9] \). Therefore,
\[ \int_R \rho \, dV = \int_2^5 \int_4^9 \frac{v}{5 + 3u} - \frac{v}{(5 + 3u)^2} \, dv \, du = \]

\[ \int_2^5 \int_4^9 \left( \frac{v}{5 + 3u} \right) \left( \frac{v}{(5 + 3u)^2} \right) \, dv \, du = \frac{4123}{19360}. \]

4. A solid, \( R \) is determined by \( 2x + 2y = 1 \), \( 2x + 2y = 10 \), \( y = 4x \), and \( y = 5x \) and the density is \( \rho = x + 1 \). Find the total mass of \( R \).

Answer:

Let \( u = \frac{y}{x} \), \( v = 2x + 2y \). Then solving these equations for \( x \) and \( y \) yields

\[ \begin{cases} x = \frac{1}{2} \frac{v}{1 + u}, \\ y = \frac{1}{2} \frac{u}{1 + u} \end{cases}. \]

Now

\[ \frac{\partial (x, y)}{\partial (u, v)} = \det \left( \begin{array}{cc} -\frac{1}{2} \frac{u}{(1+u)^2} & \frac{1}{2+2u} \\ \frac{1}{2} \frac{u}{v(1+u)} & \frac{1}{2} \frac{u}{2+u} \end{array} \right) = -\frac{1}{4} \frac{v}{(1+u)^2}. \]

Also, \( u \in [4, 5] \) while \( v \in [1, 10] \). Therefore,

\[ \int_R \rho \, dV = \int_4^5 \int_1^{10} (x + 1) \left( \frac{1}{4} \frac{v}{(1+u)^2} \right) \, dv \, du = \frac{-4 \ln 2 + 4 \ln 3}{4}. \]

5. A solid, \( R \) is determined by \( 4x + 2y = 1 \), \( 4x + 2y = 9 \), \( y = x \), and \( y = 6x \) and the density is \( \rho = y^{-1} \). Find the total mass of \( R \).

Answer:

Let \( u = \frac{y}{x} \), \( v = 4x + 2y \). Then solving these equations for \( x \) and \( y \) yields

\[ \begin{cases} x = \frac{1}{2} \frac{v}{2 + u}, \\ y = \frac{1}{2} \frac{u}{2 + u} \end{cases}. \]

Now

\[ \frac{\partial (x, y)}{\partial (u, v)} = \det \left( \begin{array}{cc} -\frac{1}{2} \frac{v}{(2+u)^2} & \frac{1}{4+2u} \\ \frac{1}{2} \frac{u}{v(2+u)} & \frac{1}{2} \frac{u}{2+2u} \end{array} \right) = -\frac{1}{4} \frac{v}{(2+u)^2}. \]

Also, \( u \in [1, 6] \) while \( v \in [1, 9] \). Therefore,

\[ \int_R \rho \, dV = \int_1^6 \int_1^9 \left( \frac{1}{2} \frac{v}{2 + u} \right)^{-1} \left( \frac{1}{4} \frac{v}{(2+u)^2} \right) \, dv \, du = -4 \ln 2 + 4 \ln 3. \]
6. Find the volume of the region, \( E \), bounded by the ellipsoid, \( \frac{1}{4}x^2 + \frac{1}{5}y^2 + \frac{1}{49}z^2 = 1 \).

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,

\( Q \equiv [0, 1] \times [0, 1] \times [0, 1] \)

onto \( R \) and

\[
\frac{\partial (x, y, z)}{\partial (u, v, w)} = \det \begin{pmatrix}
4 & 5 & 3 \\
1 & 0 & 1 \\
4 & 4 & 5
\end{pmatrix} = |-9| = 9
\]

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,

\( Q \equiv [0, 1] \times [0, 1] \times [0, 1] \)

onto \( R \) and
\[ \frac{\partial (x, y, z)}{\partial (u, v, w)} = \left| \begin{array}{ccc} 3 & 4 & 2 \\ 2 & 1 & 2 \\ 6 & 6 & 7 \end{array} \right| = | -11 | = 11 \]

and so the mass is

\[ \int_R x \, dV = \int_Q (2u + v + 2w) \, (11) \, dV \]
\[ = \int_0^1 \int_0^1 \int_0^1 (2u + v + 2w) \, (11) \, du \, dv \, dw = \frac{55}{2}. \]

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, \(Q \equiv [0, 1] \times [0, 1] \times [0, 1]\) onto \(R\) and

\[ \frac{\partial (x, y, z)}{\partial (u, v, w)} = \left| \begin{array}{ccc} 2 & 3 & 1 \\ 2 & 1 & 2 \\ 4 & 4 & 5 \end{array} \right| = | -8 | = 8 \]

and so the density is \(4u + 4v + 3w\)

\[ \int_R x \, dV = \int_Q (4u + 4v + 3w) \, (8) \, dV \]
\[ = \int_0^1 \int_0^1 \int_0^1 (4u + 4v + 3w) \, (8) \, du \, dv \, dw = 44. \]

10. Let \(D = \{(x, y) : x^2 + y^2 \leq 25\}\). Find \(\int_D e^{36x^2 + 36y^2} \, dxdy.\)

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,

\[ R = \{(r, \theta) : r \in [0, 5] \times [0, 2\pi]\}. \]

Therefore,

\[ \int_D e^{36x^2 + 36y^2} \, dV = \int_R e^{36r^2} \, r \, dV = \int_0^5 \int_0^{2\pi} e^{36r^2} r \, d\theta \, dr = \frac{1}{36} \pi \left( e^{900} - 1 \right). \]

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 $\rho \in [0, 8], \phi \in [0, \frac{1}{4} \pi], \theta \in [0, 2\pi]$. 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 $\int_0^8 \int_0^{\frac{1}{4} \pi} \int_0^{2\pi} \rho^2 \sin \phi \, d\phi \, d\theta \, d\rho = \frac{512}{3} \sqrt{2} \pi + \frac{1024}{3} \pi$.

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

$$h - r^2 \geq z \geq r, \quad r \in \left[0, \frac{1}{2} \sqrt{21} - \frac{1}{2}\right], \theta \in [0, 2\pi].$$

Also, $\frac{\partial(x,y,z)}{\partial(r,\theta,z)} = r$. Therefore, the volume is

$$\int_0^{2\pi} \int_0^{\frac{1}{2} \sqrt{21} - \frac{1}{2}} \int_0^{r^2} r \, dz \, dr \, d\theta = \frac{39}{4} \pi + \frac{1}{4} \pi \sqrt{21}.$$

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

$$\int_0^4 \int_0^{2\pi} \int_{-\sqrt{(144-r^2)}}^{\sqrt{(144-r^2)}} r \, dz \, d\theta \, dr = -\frac{4096}{3} \sqrt{2} \pi + 2304 \pi.$$

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:

\[
\int_0^{2\pi} \int_0^{\arcsin\left(\frac{2}{\sqrt{30}}\right)} \int_0^{\sec \phi} \rho^3 \sin \phi \, d\rho \, d\phi \, d\theta + \int_0^{\arcsin\left(\frac{2}{\sqrt{30}}\right)} \int_0^{11} \rho^3 \sin \phi \, d\rho \, d\phi \, d\theta
\]

\[
= \frac{24623}{3} \pi
\]

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 \( \int_S \left(\frac{x}{y}\right)^2 + 1 \) \( dV \) 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 \( 0 \leq \theta \leq \frac{1}{6} \pi \).

Answer:

\[
\int_0^{\pi/6} \int_0^2 \sin \left(4r^2\right) r \, dr \, d\theta = -\frac{1}{48} \pi \cos 16 + \frac{1}{48} \pi
\]

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

\[
\int_0^{\pi/4} \int_0^2 re^{2r^2} \, dr \, d\theta = \frac{1}{12} \pi \left(e^8 - 1\right).
\]

20. Compute the volume of a sphere of radius \( R \) using cylindrical coordinates.

Answer:

Using cylindrical coordinates, the integral is

\[
\int_0^{2\pi} \int_0^R \int_{-\sqrt{R^2-r^2}}^{\sqrt{R^2-r^2}} r \, dz \, dr \, d\theta = \frac{4}{3} \pi R^3.
\]
22.5  The Moment Of Inertia

In order to appreciate the importance of this concept, it is necessary to discuss its physical significance.

22.5.1 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, \( \mathbf{r}_i(t) \) for \( i = 1, 2, \cdots, N \) 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 [24].

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 14.4.2 on Page 301, there exists an angular velocity vector \( \mathbf{\Omega}(t) \) such that if \( \mathbf{u}(t) \) is the position vector of a point fixed in the top, \( \mathbf{u}(t) = u_1 \mathbf{i}(t) + u_2 \mathbf{j}(t) + u_3 \mathbf{k}(t) \),

\[
\mathbf{u}'(t) = \mathbf{\Omega}(t) \times \mathbf{u}(t).
\]

The vector \( \mathbf{\Omega}_a \) shown in the picture is the vector for which

\[
\mathbf{r}_i'(t) = \mathbf{\Omega}_a \times \mathbf{r}_i(t)
\]

is the velocity of the \( i^{th} \) point mass due to rotation about the axis of the top. Thus \( \mathbf{\Omega}(t) = \mathbf{\Omega}_a(t) + \mathbf{\Omega}_p(t) \) and it is assumed \( \mathbf{\Omega}_p(t) \) is very small relative to \( \mathbf{\Omega}_a \). 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, \( L \) is
defined by

\[
L \equiv \sum_{i=1}^{N} r_i \times m_i \mathbf{v}_i \tag{22.3}
\]

where \( \mathbf{v}_i \) equals the velocity of the \( i \)th point mass. Thus \( \mathbf{v}_i = \Omega(t) \times r_i \) and from the above assumption, \( \mathbf{v}_i \) may be taken equal to \( \Omega_a \times r_i \). Therefore, \( L \) is essentially given by

\[
L \equiv \sum_{i=1}^{N} m_i r_i \times (\Omega_a \times r_i)
\]

By symmetry of the top, this last expression equals a multiple of \( \Omega_a \). Thus \( L \) is parallel to \( \Omega_a \). Also,

\[
L \cdot \Omega_a = \sum_{i=1}^{N} m_i |r_i|^2 \Omega_a - (r_i \cdot \Omega_a) r_i.
\]

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 \( L \) has the same direction as \( \Omega_a \).

Let \( \omega \equiv |\Omega_a| \). Then the above expression is of the form

\[
L \cdot \Omega_a = I \omega^2,
\]

where

\[
I \equiv \sum_{i=1}^{N} m_i |r_i|^2 \sin^2 (\beta_i).
\]

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

\[
L = I \omega \mathbf{u}. \tag{22.4}
\]

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

\[
(\mathbf{u} \cdot \mathbf{i}) = (\mathbf{y} \cdot \mathbf{i}) = \sin \alpha \cos \theta. \tag{22.5}
\]
Now also from (22.3),
\[
\frac{dL}{dt} = \sum_{i=1}^{N} m_i \mathbf{r}'_i \times \mathbf{v}'_i + \mathbf{r}_i \times m_i \mathbf{v}'_i
\]
\[
= \sum_{i=1}^{N} \mathbf{r}_i \times m_i \mathbf{v}'_i = -\sum_{i=1}^{N} \mathbf{r}_i \times m_i g \mathbf{k}
\]
where \( g \) is the acceleration of gravity. From (22.4), (22.5), and the above,
\[
\frac{dL}{dt} \cdot \mathbf{i} = I\omega \left( \frac{du}{dt} \cdot \mathbf{i} \right) = I\omega \left( \frac{dy}{dt} \cdot \mathbf{i} \right)
\]
\[
= (-I\omega \sin \alpha \sin \theta) \theta' = -\sum_{i=1}^{N} \mathbf{r}_i \times m_i g \mathbf{k} \cdot \mathbf{i}
\]
\[
= -\sum_{i=1}^{N} m_i g \mathbf{r}_i \cdot \mathbf{k} \times \mathbf{i} = -\sum_{i=1}^{N} m_i g \mathbf{r}_i \cdot \mathbf{j}. \tag{22.6}
\]
To simplify this further, recall the following definition of the center of mass.

**Definition 22.5.1** Define the total mass, \( M \) by
\[
M = \sum_{i=1}^{N} m_i
\]
and the center of mass, \( \mathbf{r}_0 \) by
\[
\mathbf{r}_0 \equiv \frac{\sum_{i=1}^{N} \mathbf{r}_i m_i}{M}. \tag{22.7}
\]

In terms of the center of mass, the last expression equals
\[
-Mg \mathbf{r}_0 \cdot \mathbf{j} = -Mg \left( \mathbf{r}_0 - (\mathbf{r}_0 \cdot \mathbf{k}) \mathbf{k} + (\mathbf{r}_0 \cdot \mathbf{k}) \mathbf{k} \right) \cdot \mathbf{j}
\]
\[
= -Mg \left( \mathbf{r}_0 - (\mathbf{r}_0 \cdot \mathbf{k}) \mathbf{k} \right) \cdot \mathbf{j}
\]
\[
= -Mg |\mathbf{r}_0 - (\mathbf{r}_0 \cdot \mathbf{k}) | \cos \theta
\]
\[
= -Mg |\mathbf{r}_0| \sin \alpha \cos \left( \frac{\pi}{2} - \theta \right).
\]
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 \( \Omega_a \), and also \(|\mathbf{r}_0|\) is independent of \( t \). Therefore, from the second line of (22.4),
\[
(-I\omega \sin \alpha \sin \theta) \theta' = -Mg |\mathbf{r}_0| \sin \alpha \sin \theta.
\]
which shows
\[
\theta' = \frac{Mg |\mathbf{r}_0|}{I\omega}. \tag{22.8}
\]
From (22.3), 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 \( \Omega_a \) is approximately equal to \( \Omega \). It also leaves out all considerations of friction and the observation that the axis of symmetry wobbles. This is wobbling is called **nutation**. 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.
22.5.2 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 \( r_i' = v_i \). The kinetic energy of the mass \( m_i \) is defined by

\[
\frac{1}{2} m_i |r_i'|^2.
\]

(This is a very good time to review the presentation on kinetic energy given on Page 277.)

The total kinetic energy of the collection of masses is then

\[
E = \sum_{i=1}^{N} \frac{1}{2} m_i |r_i'|^2.
\]  \hspace{1cm} (22.9)

As these masses move about, so does the center of mass, \( r_0 \). Thus \( r_0 \) is a function of \( t \) just as the other \( r_i \). From (22.9) the total kinetic energy is

\[
E = \sum_{i=1}^{N} \frac{1}{2} m_i \left[ |r_i' - r_0'|^2 + |r_0'|^2 + 2 (r_i' - r_0' \cdot r_0') \right].
\]  \hspace{1cm} (22.10)

Now

\[
\sum_{i=1}^{N} m_i (r_i' - r_0') = \left( \sum_{i=1}^{N} m_i (r_i - r_0) \right)' = \left( \sum_{i=1}^{N} m_i r_i \right)' \cdot r_0' = 0
\]

because from (22.8)

\[
\sum_{i=1}^{N} m_i (r_i - r_0) = \sum_{i=1}^{N} m_i r_i - \sum_{i=1}^{N} m_i r_0 = \sum_{i=1}^{N} m_i r_i - \sum_{i=1}^{N} m_i \left( \frac{\sum_{i=1}^{N} r_i m_i}{\sum_{i=1}^{N} m_i} \right) = 0.
\]

Let \( M \equiv \sum_{i=1}^{N} m_i \) be the total mass. Then (22.10) reduces to

\[
E = \sum_{i=1}^{N} \frac{1}{2} m_i \left[ |r_i' - r_0'|^2 + |r_0'|^2 \right]
\[
= \frac{1}{2} M |r_0'|^2 + \sum_{i=1}^{N} \frac{1}{2} m_i |r_i' - r_0'|^2.
\]  \hspace{1cm} (22.11)

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 one 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 [14.4.2] on Page [301] 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 (22.11),

$$E = \frac{1}{2} M |r'_0|^2 + \sum_{i=1}^{N} \frac{1}{2} m_i |\Omega(t) \times (r_i(t) - r_0(t))|^2$$

$$= \frac{1}{2} M |r'_0|^2 + \frac{1}{2} \left( \sum_{i=1}^{N} m_i |r_i(t) - r_0(t)|^2 \sin^2 \theta_i \right) |\Omega(t)|^2$$

$$= \frac{1}{2} M |r'_0|^2 + \frac{1}{2} \left( \sum_{i=1}^{N} m_i |r_i(0) - r_0(0)|^2 \sin^2 \theta_i \right) |\Omega(t)|^2$$

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.

### 22.6 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. This leads to the following definition.

**Definition 22.6.1** Let a solid occupy a region $R$ such that its density is $\delta(x)$ for
$x$ a point in $R$ and let $L$ be a line. For $x \in R,$ let $l(x)$ be the distance from the point, $x$ to
the line $L.$ The moment of inertia of the solid is defined as

$$\int_{R} l(x)^2 \delta(x) dV.$$
Letting \((x_c, y_c, z_c)\) denote the Cartesian coordinates of the center of mass,

\[
x_c = \frac{\int_R x \delta(x) \, dV}{\int_R \delta(x) \, dV}, \quad y_c = \frac{\int_R y \delta(x) \, dV}{\int_R \delta(x) \, dV}, \quad z_c = \frac{\int_R z \delta(x) \, dV}{\int_R \delta(x) \, dV},
\]

where \(x, y, z\) are the Cartesian coordinates of the point at \(x\).

**Example 22.6.2** 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.

\[
\int_R (x^2 + y^2) \rho(x) \, dV. \tag{22.12}
\]

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

\[
\left( \frac{\int_R x r \rho \, dV}{\int_R r \rho \, dV}, \frac{\int_R y r \rho \, dV}{\int_R r \rho \, dV}, \frac{\int_R z r \rho \, dV}{\int_R r \rho \, dV} \right) = \frac{\int_R r \rho \, dV}{\int_R r \rho \, dV}.
\]

Here is a specific example.

**Example 22.6.3** 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

\[
\int_0^{2\pi} \int_0^3 \int_0^{9-r^2} r^3 r \, dz \, dr \, d\theta = \frac{8748}{35}\pi.
\]

To find the center of mass, note the \(x\) and \(y\) coordinates of the center of mass,

\[
\frac{\int_R x r \rho \, dV}{\int_R r \rho \, dV}, \quad \frac{\int_R y r \rho \, dV}{\int_R r \rho \, dV},
\]

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, \(z_c\). In polar coordinates, \(\rho = r\) and so,

\[
z_c = \frac{\int_R z r \rho \, dV}{\int_R r \rho \, dV} = \frac{\int_0^{2\pi} \int_0^3 \int_0^{9-r^2} z r^2 \, dz \, dr \, d\theta}{\int_0^{2\pi} \int_0^3 \int_0^{9-r^2} r^2 \, dz \, dr \, d\theta} = \frac{18}{7}.
\]
Thus the center of mass will be \((0, 0, \frac{18}{7})\).

A short comment about terminology is in order. When the density is constant, the center of mass is called the centroid. Thus the centroid is a purely geometrical concept because the densities will cancel from the integrals.

### 22.7 Exercises With Answers

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.

The region, \(R\) is a dome shaped region above the circle centered at the origin having radius 2. Therefore, using polar or cylindrical coordinates

\[
z_c = \frac{\int_R z \sigma \, dV}{\int_R \sigma \, dV} = \frac{\int_0^{2\pi} \int_0^2 \int_0^{4-r^2} z r \, dz \, dr \, d\theta}{\int_0^{2\pi} \int_0^2 \int_0^{4-r^2} r \, dz \, dr \, d\theta} = \frac{4}{3}
\]

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, \(\sigma\) is equals \(\sigma(x, y, z) = z\).

This problem is just like the one above except here the density is not constant. Thus

\[
z_c = \frac{\int_R z^2 \sigma \, dV}{\int_R \sigma \, dV} = \frac{\int_0^{2\pi} \int_0^2 \int_0^{4-r^2} z^2 r \, dz \, dr \, d\theta}{\int_0^{2\pi} \int_0^2 \int_0^{4-r^2} r \, dz \, dr \, d\theta} = 2
\]

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 \(\sigma = 1\).

To find where \((x, y)\) is you let \(-y^2 + 8 = 2x^2 + y^2\) and this shows the two surfaces intersect in the circle \(x^2 + y^2 = 4\). Using cylindrical coordinates,

\[
z_c = \frac{\int_R z \, dV}{\int_R dV} = \frac{\int_0^{2\pi} \int_0^2 \int_0^{8-r^2 \sin^2(\theta)} z r \, dz \, dr \, d\theta}{\int_0^{2\pi} \int_0^2 \int_0^{8-r^2 \sin^2(\theta)} r \, dz \, dr \, d\theta} = \frac{14}{3}
\]

You can find the other the same way.

\[
x_c = \frac{\int_R x \, dV}{\int_R dV} = \frac{\int_0^{2\pi} \int_0^2 \int_0^{8-r^2 \sin^2(\theta)} (r \, \cos(\theta)) r \, dz \, dr \, d\theta}{\int_0^{2\pi} \int_0^2 \int_0^{8-r^2 \sin^2(\theta)} r \, dz \, dr \, d\theta} = 0
\]

\[
y_c = \frac{\int_R y \, dV}{\int_R dV} = \frac{\int_0^{2\pi} \int_0^2 \int_0^{8-r^2 \sin^2(\theta)} (r \, \sin(\theta)) r \, dz \, dr \, d\theta}{\int_0^{2\pi} \int_0^2 \int_0^{8-r^2 \sin^2(\theta)} r \, dz \, dr \, d\theta} = 0
\]

Thus the center of mass is \(\left(0, 0, \frac{14}{3}\right)\). The mass is

\[
\int_0^{2\pi} \int_0^2 \int_0^{8-r^2 \sin^2(\theta)} r \, dz \, dr \, d\theta = 16\pi
\]
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 $\sigma(x, y, z) = x^2$.

This is just like the problem above only now the density is not constant.

The mass is

$$z_c = \frac{\int_R \int_{-2}^{2} \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} z \left(r^2 \cos^2(\theta)\right) r \, dz \, dr \, d\theta}{\int_R \int_{-2}^{2} \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} r \left(r^2 \cos^2(\theta)\right) r \, dz \, dr \, d\theta} = \frac{11}{2}$$

$$y_c = \frac{\int_R \int_{-2}^{2} \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} \left(r \sin(\theta)\right) r \left(r^2 \cos^2(\theta)\right) r \, dz \, dr \, d\theta}{\int_R \int_{-2}^{2} \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} r \left(r^2 \cos^2(\theta)\right) r \, dz \, dr \, d\theta} = 0$$

$$x_c = \frac{\int_R \int_{-2}^{2} \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} \left(r \cos(\theta)\right) r \left(r^2 \cos^2(\theta)\right) r \, dz \, dr \, d\theta}{\int_R \int_{-2}^{2} \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} r \left(r^2 \cos^2(\theta)\right) r \, dz \, dr \, d\theta} = 0$$

So in this case the center of mass is $(0, 0, \frac{11}{2})$. The mass is

$$\int_0^{2\pi} \int_0^2 \int_{2r^2 \cos^2(\theta) + r^2 \sin^2(\theta)}^{8 - r^2 \sin^2(\theta)} r \left(r^2 \cos^2(\theta)\right) r \, dz \, dr \, d\theta = \frac{32}{3} \pi$$

5. The two cylinders, $x^2 + y^2 = 4$ and $y^2 + z^2 = 4$ intersect in a region, $R$. Find the mass and center of mass if the density, $\sigma$, is given by $\sigma(x, y, z) = z^2$.

The first cylinder is parallel to the $z$ axis. Let $D$ denote the circle of radius 2 in the $xy$ plane. Then the region just described has $(x, y)$ in the circle of radius 2 and $z$ between $-\sqrt{4 - y^2}$ and $\sqrt{4 - y^2}$. It follows the total mass is

$$\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} z^2 \, dz \, dx \, dy = \frac{2048}{45}.$$  

By symmetry, the center of mass will be $(0, 0, 0)$.

6. The two cylinders, $x^2 + y^2 = 4$ and $y^2 + z^2 = 4$ intersect in a region, $R$. Find the mass and center of mass if the density, $\sigma$, is given by $\sigma(x, y, z) = 4 + z$.

The total mass is

$$\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} (4 + z) \, dz \, dx \, dy = \frac{512}{3}$$

$$z_c = \frac{\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} z (4 + z) \, dz \, dx \, dy}{\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} (4 + z) \, dz \, dx \, dy} = \frac{4}{15}$$

$$x_c = \frac{\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} x (4 + z) \, dz \, dx \, dy}{\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} (4 + z) \, dz \, dx \, dy} = 0$$

$$y_c = \frac{\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} y (4 + z) \, dz \, dx \, dy}{\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} (4 + z) \, dz \, dx \, dy} = 0$$

and so the center of mass is $(0, 0, \frac{4}{15})$. 


7. Find the mass and center of mass of the set, \((x, y, z)\) such that \(\frac{x^2}{4} + \frac{y^2}{9} + z^2 \leq 1\) if the density is \(\sigma(x, y, z) = 4 + y + z\).

This is the inside of an ellipsoid. Denote this by \(R\). Then the total mass is

\[
\int_R \sigma \, dV = \int_R (4 + y + z) \, dV
\]

Lets change the variables. Let \(x = 2u, y = 3v, z = w\). When this is done, \((u, v, w)\) will be in the unit ball. The Jacobian of this transformation is 6. Now changing the variables the above integral equals

\[
\int_B (4 + 3v + w) \, 6 \, dV
\]

where here \(B\) is the unit ball. When integrating over a ball, you ought to suspect that spherical coordinates would be a good idea. Change the variables again in the above integral to spherical coordinates.

\[
w = \rho \cos \phi, v = \rho \sin \phi \sin \theta, u = \rho \sin \phi \cos \theta.
\]

Then the above integral in spherical coordinates is

\[
\int_0^\pi \int_0^{2\pi} \int_0^1 (4 + 3\rho \sin(\phi) \sin(\theta) + \rho \cos \phi) \, 6\rho^2 \sin \phi \, d\rho \, d\theta \, d\phi = 32\pi.
\]

To find the center of mass, it would be

\[
z_c = \frac{\int_R z(4 + y + z) \, dV}{32\pi} = \frac{\int_R z^2 \, dV}{32\pi}.
\]

Now to get this, I have used symmetry of the region. This equals

\[
\frac{\int_R w^2 \, dV}{32\pi} = \frac{\int_0^\pi \int_0^{2\pi} \int_0^1 (\rho \cos(\phi))^2 \, 6\rho^2 \sin \phi \, d\rho \, d\theta \, d\phi}{32\pi} = \frac{1}{20}
\]

\[
y_c = \frac{\int_R y(4 + y + z) \, dV}{32\pi} = \frac{\int_R y^2 \, dV}{32\pi}
\]

\[
= \frac{\int_0^\pi \int_0^{2\pi} \int_0^1 (3(\rho \sin(\phi) \sin(\theta)))^2 \, 6\rho^2 \sin \phi \, d\rho \, d\theta \, d\phi}{32\pi} = \frac{9}{20}
\]

I think you get the idea. You can now find \(x_c\) in the same way.

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 given the density equals 1.

Using cylindrical coordinates, this is

\[
\int_R r^2 \, dV = \int_0^{2\pi} \int_0^3 \int_0^{9-r^2} r^2 \, dz \, dr \, d\theta = \frac{243}{2} \pi
\]

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 given the density equals 1.

It is like the above except different.

\[
\int_0^{2\pi} \int_0^3 \int_0^{9-r^2} (r^2 \sin^2(\theta) + z^2) \, r \, dz \, dr \, d\theta = \frac{1215}{2} \pi
\]
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 \( \frac{2}{5} R^2 M \) where \( M \) is the mass.

The constant density of the ball is \( \frac{3}{4} \frac{M}{\pi R^3} \). For simplicity let the line be the \( z \) axis. I will also use spherical coordinates since this is a ball. Then the moment of inertia is

\[
\int_0^\pi \int_0^{2\pi} \int_0^R \frac{3}{4} \frac{M}{\pi R^3} (\rho \sin (\phi))^2 \rho^2 \sin (\phi) \, d\rho d\theta d\phi = \frac{2}{5} R^2 M.
\]

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.

\[
\int_0^\pi \int_0^{2\pi} \int_0^R \rho (\rho \sin (\phi))^2 \rho^2 \sin (\phi) \, d\rho d\theta d\phi = \frac{4}{9} \pi R^6
\]

Also the total mass is

\[
M = \int_0^\pi \int_0^{2\pi} \int_0^R \rho \rho^2 \sin (\phi) \, d\rho d\theta d\phi = \pi R^4
\]

Therefore, the moment of inertia is

\[
\frac{4}{9} MR^2.
\]

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 \( \frac{1}{2} R^2 M \) where \( M \) is the mass.

The density is \( \frac{M}{\pi R^2 h} \) where \( h \) is the height of the cylinder. Using cylindrical coordinates, the moment of inertia is

\[
\int_0^{2\pi} \int_0^R \int_0^h \left( \frac{M}{\pi R^2 h} \right) r^2 r dr dz d\theta = \frac{1}{2} R^2 M
\]

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?

The sphere will win. This is because it takes less torque to produce a given angular acceleration in the sphere than in the cylinder because the moment of inertia for the sphere is less than the moment of inertia of the cylinder. Thus a given torque about the axis of rotation, which will be identical in both will produce faster rotation in the sphere than in the cylinder. Another way to look at it is that they both have the same total energy when they get to the bottom. This energy comes from two parts, one involving rotation and the other translation of the center of mass. If the center of mass of both were moving at the same speed, this would be a contradiction because the different moments of inertia would then require the kinetic energy of one to be greater than that of the other.
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.

Consider the following picture in which, as suggested, the line, $l$ is the $z$ axis and $l_1$ goes through $(a,0)$ in the $xy$ plane and is parallel to $l$.

For a point in $B$, let the coordinates of this point be $(x, y, z)$. Then the displacement vector from a point, $(a,0,z)$ on $l_1$ to the point, $(x,y,z)$ is $(x-a,-y,0)$ and so the square of the distance is $x^2 - 2xa + a^2 + y^2$. Therefore, from the definition of moment of inertia, the moment of inertia about $l_1$ is

$$I_{l_1} = \int_B \delta(x,y,z) \left( x^2 - 2xa + a^2 + y^2 \right) dV$$

Since the line goes through the center of mass, this reduces to

$$\int_B \delta(x,y,z) \left( x^2 + y^2 \right) dV + \int_B \delta(x,y,z) a^2 dV = I_l + a^2 M$$

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?

The answer is yes. To see this, consider the parallel axis theorem above.
CHAPTER 22. THE INTEGRAL IN OTHER COORDINATES
Part VII

Line Integrals
Chapter 23

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.

23.0.1 Orientations And Smooth Curves

Recall the notion of a smooth curve.

A smooth curve in \( \mathbb{R}^n \) if there exists an interval, \([a, b] \subseteq \mathbb{R}\) and functions \(x_i : [a, b] \to \mathbb{R}\) such that the following conditions hold

1. \(x_i\) is continuous on \([a, b]\).
2. \(x_i'(a)\) 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 \(p(t) \equiv (x_1(t), \ldots, x_n(t))\), \(t \to p(t)\) is one to one on \((a, b)\).
4. \(|p'(t)| \equiv \left(\sum_{i=1}^{n} |x_i'(t)|^2\right)^{1/2} \neq 0\) for all \(t \in [a, b]\).
5. \(C = \cup \{(x_1(t), \ldots, x_n(t)) : t \in [a, b]\}\).

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 **parametrization** 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 proof that curve length is well defined for a smooth curve contains a result which deserves to be stated as a corollary. It is proves in the Section which starts on Page 473. This is one of those sections you should read only if you are interested.

**Corollary 23.0.1** Let \(C\) be a smooth curve and let \(f : [a, b] \to C\) and \(g : [c, d] \to C\) be two parameterizations satisfying \(\mathbb{R}^n\). Then \(g^{-1} \circ f\) is either strictly increasing or strictly decreasing.

473
Definition 23.0.2 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 based only on the assumption that the parameterizations of the curve are one to one.

Lemma 23.0.3 The following hold for \( \sim \).

\[
\begin{align*}
f \sim f, & \quad (23.1) \\
\text{If } f \sim g \text{ then } g \sim f, & \quad (23.2) \\
\text{If } f \sim g \text{ and } g \sim h, & \quad (23.3)
\end{align*}
\]

Proof: Formula 23.1 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 23.2. To see 23.3, \( f^{-1} \circ h = (f^{-1} \circ g) \circ (g^{-1} \circ h) \) 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 parametrization of \( C \), consider \( g(t) \equiv f \left((a + b) - t\right) \), also a parametrization of \( C \). Now by Corollary 23.0.1, if \( h \) is a parametrization, 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 23.0.4 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! If it were a banister and you were sliding down it, it would be clear at a certain point that the curve is not smooth. I think you may even get the point of this from the picture below.

So what is the thing to remember from all this? First, there are certain conditions which must be satisfied for a curve to be smooth. These are listed in \( \square \). Next, if you have any curve, there are two directions you can move over this curve, each called an orientation. This is illustrated in the following picture.
Either you move from \( p \) to \( q \) or you move from \( q \) to \( p \).

**Definition 23.0.5** A curve \( C \) is piecewise smooth if there exist points on this curve, \( p_0, p_1, \ldots, p_n \) such that, denoting \( C_{p_{k-1}p_k} \) the part of the curve joining \( p_{k-1} \) and \( p_k \), it follows \( C_{p_{k-1}p_k} \) is a smooth curve and \( \bigcup_{k=1}^n C_{p_{k-1}p_k} = C \). In other words, it is piecewise smooth if it consists of a finite number of smooth curves linked together.

Note that Example 23.0.4 is an example of a piecewise smooth curve although it is not smooth.

23.0.2 The Integral Of A Function Defined On A Smooth Curve

Letting \( r(t), t \in [a,b] \) be the position vector of a smooth curve, recall that the total length of this curve is given by

\[
l = \int_a^b |r'(t)| \, dt. \tag{23.4}
\]

Remember that if you interpret \( t \) as time, \( |r'(t)| \) is the speed and the above integral says that to get the total distance you simply integrate the speed. A small chunk of distance traveled is \( dl = |r'(t)| \, dt \). This says the same thing as

\[
\frac{dl}{dt} = |r'(t)|
\]

which was discussed earlier. Of course it follows from (23.3) and the fundamental theorem of calculus. The distance for the parameter between \( a \) and \( t \) is

\[
l(t) = \int_a^t |r'(s)| \, ds
\]

and so by the fundamental theorem of calculus,

\[
l'(t) = \frac{dl}{dt} = |r'(t)|.
\]

For this reason, the increment of arc length is \( dl = |r'(t)| \, dt \). Think of it as giving an infinitesimal contribution to the integral. For \( C \) a smooth curve with a parametrization, \( r : [a, b] \to C \) and a function, \( f \) defined on \( C \), define the symbol,

\[
\int_C f \, dl = \int_a^b f(r(t)) |r'(t)| \, dt.
\]

**Example 23.0.6** Let \( C \) be a smooth curve which has parametrization given by \( r(t) = (\cos 2t, \sin(2t), t) \) for \( t \in [0, 2\pi] \). Suppose \( f(x, y, z) = x^2 + y \). Find \( \int_C f \, dl \).
The increment of length is \( \sqrt{4 \cos^2(2t) + 4 \sin^2(2t) + 1} \,
dt \). Now the desired integral is
\[
\int_0^{2\pi} (\cos^2 2t + \sin 2t) \sqrt{5} \dt = \sqrt{5}\pi
\]

One can define things like density with respect to arc length in the usual way. As just explained, a little chunk of length is \( dl = |r'(t)| \dt \). The density is a function \( \delta(x,y,z) \) which has the property that a little chunk of mass is given by \( dm = \delta(r(t)) \, dl \).

**Definition 23.0.7** Let \( \delta \) be the density with respect to arc length. Then the total mass of a smooth curve, \( C \) having parametrization \( r: [a,b] \rightarrow \mathbb{R}^3 \) is
\[
\int_C \delta(x,y,z) \, dl = \int_a^b \delta(r(t)) \, |r'(t)| \, dt
\]

the center of mass can be given in a similar manner as before. Thus
\[
x_c = \frac{\int_C x \delta(x,y,z) \, dl}{\int_C \delta(x,y,z) \, dl},
\]
\[
y_c = \frac{\int_C y \delta(x,y,z) \, dl}{\int_C \delta(x,y,z) \, dl},
\]
\[
z_c = \frac{\int_C z \delta(x,y,z) \, dl}{\int_C \delta(x,y,z) \, dl}
\]

and the only thing you need to do is to evaluate the integrals after changing everything to give a one dimensional integral with respect to the parameter \( t \).

**Example 23.0.8** Let a smooth curve be given by the parametrization, \( r(t) = (\cos t, \sin t, t) : t \in [0,10] \). This is a helix in case you are interested. Suppose the density is given by \( \delta(x,y,z) = x^2 \). Find the total mass and the center of mass.

The increment of arc length is \( dl = \sqrt{\sin^2(t) + \cos^2(t) + 1} \dt = \sqrt{2} \dt \). Then the total mass is
\[
\int_0^{10} \cos^2(t) \sqrt{2} \dt = \frac{1}{2} \sqrt{2} \cos(10) \sin(10) + 5\sqrt{2}
\]

The center of mass is given by
\[
x_c = \frac{\int_0^{10} (\cos(t)) \cos^2(t) \sqrt{2} \dt}{\frac{1}{2} \sqrt{2} \cos(10) \sin(10) + 5\sqrt{2}} = \frac{\frac{1}{2} \sqrt{2} \sin 10 \cos^2 10 + \frac{3}{2} \sqrt{2} \sin 10}{\frac{1}{2} \sqrt{2} \cos 10 \sin 10 + 5\sqrt{2}}
\]
\[
y_c = \frac{\int_0^{10} (\sin(t)) \cos^2(t) \sqrt{2} \dt}{\frac{1}{2} \sqrt{2} \cos(10) \sin(10) + 5\sqrt{2}} = \frac{-\frac{1}{3} (\cos^3 10) \sqrt{2} + \frac{1}{3} \sqrt{2}}{\frac{1}{2} \sqrt{2} \cos 10 \sin 10 + 5\sqrt{2}}
\]
\[
z_c = \frac{\int_0^{10} t \cos^2(t) \sqrt{2} \dt}{\frac{1}{2} \sqrt{2} \cos(10) \sin(10) + 5\sqrt{2}} = \frac{5\sqrt{2} \cos 10 \sin 10 + \frac{99}{4} \sqrt{2} + \frac{1}{4} \sqrt{2} \cos^2 10}{\frac{1}{2} \sqrt{2} \cos 10 \sin 10 + 5\sqrt{2}}
\]

**23.0.3 Vector Fields**

A **vector field** is nothing but a function which has values which are vectors. For example, consider the force acting on a unit mass by the sun. This determines a force vector which depends on the location of the point. Thus each point in space has associated with it a vector which is the force which the sun exerts on a particle of mass 1 which is placed at that point.
Some people find it useful to try and draw pictures to illustrate a vector valued function or vector field. 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 23.0.9** Draw a picture of the vector field, $(-x, y)$ which gives the velocity of a fluid flowing in two dimensions.

![Diagram](image)

In this example, drawn by Maple, you can see how the arrows indicate the motion of this fluid.

**Example 23.0.10** Draw a picture of the vector field $(y, x)$ for the velocity of a fluid flowing in two dimensions.

![Diagram](image)

So much for art. Get the computer to do it and it can be useful. If you try to do it, you will mainly waste time.
Example 23.0.11 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.

An example of a vector field is the gradient of a scalar function. Some vector fields are gradients and some are not. When a vector field is the gradient, we call the vector field conservative.

23.0.4 Line Integrals And Work

The interesting concept of line integral has to do with integrals which involve vector fields, not scalar valued functions as above. The most significant application is to work.

First, it is necessary to give some discussion of the concept of orientation. 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 discussed in Definition 23.0.2, 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 23.0.12 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 force, $\mathbf{F}$ on an object which moves from the point $x(t)$ to the point $x(t+h)$ along the straight line shown would equal $\mathbf{F} \cdot (x(t+h) - x(t))$.

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

$$dW = \mathbf{F}(x(t)) \cdot x'(t) \, dt$$

Thus the total work along the whole curve should be given by the integral,

$$\int_a^b \mathbf{F}(x(t)) \cdot x'(t) \, dt$$

This motivates the following definition of work.

**Definition 23.0.13** Let $\mathbf{F}(x)$ 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

$$\int_C \mathbf{F} \cdot d\mathbf{R} \equiv \int_a^b \mathbf{F}(x(t)) \cdot x'(t) \, dt$$

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 23.0.14** The symbol, $\int_C \mathbf{F} \cdot d\mathbf{R}$, is well defined in the sense that every parametrization 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 parametrization. Thus $g^{-1} \circ f$ is an increasing function, $\phi$. Letting $s = \phi(t)$ and changing variables, and using the fact $\phi$ is increasing,

$$\int_c^d \mathbf{F}(g(s)) \cdot g'(s) \, ds = \int_a^b \mathbf{F}(g(\phi(t))) \cdot g'(\phi(t)) \phi'(t) \, dt$$

$$= \int_a^b \mathbf{F}(f(t)) \cdot \frac{d}{dt} \left(g(g^{-1} \circ f(t))\right) \, dt = \int_a^b \mathbf{F}(f(t)) \cdot f'(t) \, dt.$$
This proves the theorem.

Regardless the physical interpretation of $\mathbf{F}$, this is called the line integral. When $\mathbf{F}$ 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 negative work.

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 $\mathbf{p}$ and $\mathbf{q}$ be two points in $\mathbb{R}^n$ and suppose $\mathbf{F}$ is a constant force acting on an object which moves from $\mathbf{p}$ to $\mathbf{q}$ along the straight line joining these points. Then the work done is $\mathbf{F} \cdot (\mathbf{q} - \mathbf{p})$. Is the same thing obtained from the above definition? Let $\mathbf{x}(t) \equiv \mathbf{p} + t(\mathbf{q} - \mathbf{p})$, $t \in [0,1]$ be a parametrization for this oriented curve, the straight line in the direction from $\mathbf{p}$ to $\mathbf{q}$.

Then $\mathbf{x}'(t) = \mathbf{q} - \mathbf{p}$ and $\mathbf{F}(\mathbf{x}(t)) = \mathbf{F}$. Therefore, the above definition yields

$$\int_0^1 \mathbf{F} \cdot (\mathbf{q} - \mathbf{p}) \, dt = \mathbf{F} \cdot (\mathbf{q} - \mathbf{p}).$$

Therefore, the new definition adds to but does not contradict the old one.

**Example 23.0.15** Suppose for $t \in [0, \pi]$ the position of an object is given by $\mathbf{r}(t) = t\mathbf{i} + \cos (2t)\mathbf{j} + \sin (2t)\mathbf{k}$. Also suppose there is a force field defined on $\mathbb{R}^3$, $\mathbf{F}(x,y,z) \equiv 2xy\mathbf{i} + x^2\mathbf{j} + \mathbf{k}$. 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 $t$.

To find this line integral use the above definition and write

$$\int_C \mathbf{F} \cdot d\mathbf{R} = \int_0^\pi (2t (\cos (2t)), t^2, 1) \cdot (1, -2\sin (2t), 2\cos (2t)) \, dt$$

In evaluating this replace the $x$ in the formula for $\mathbf{F}$ with $t$, the $y$ in the formula for $\mathbf{F}$ with $\cos (2t)$ and the $z$ in the formula for $\mathbf{F}$ 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 23.0.16** Let $C$ denote the oriented curve obtained by $\mathbf{r}(t) = (t, \sin t, t^3)$ where the orientation is determined by increasing $t$ for $t \in [0,2]$. Also let $\mathbf{F} = (x, y, xz + z)$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$.

You use the definition.

$$\int_C \mathbf{F} \cdot d\mathbf{R} = \int_0^2 (t, \sin (t), (t + 1) t^3) \cdot (1, \cos (t), 3t^2) \, dt$$

$$= \int_0^2 (t + \sin (t) \cos (t) + 3 (t + 1) t^5) \, dt$$

$$= \int_0^2 \frac{1251}{14} - \frac{1}{2} \cos^2 (2).$$
23.0.5 Another Notation For Line Integrals

Definition 23.0.17 Let \( \mathbf{F}(x, y, z) = (P(x, y, z), Q(x, y, z), R(x, y, z)) \) and let \( C \) be an oriented curve. Then another way to write \( \int_C \mathbf{F} \cdot d\mathbf{R} \) is

\[
\int_C P\,dx + Q\,dy + R\,dz
\]

This last is referred to as the integral of a **differential form**, \( P\,dx + Q\,dy + R\,dz \). 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.

23.1 Exercises With Answers

1. Suppose for \( t \in [0, 2\pi] \) the position of an object is given by \( \mathbf{r}(t) = 2t\mathbf{i} + \cos(t)\mathbf{j} + \sin(t)\mathbf{k} \). Also suppose there is a force field defined on \( \mathbb{R}^3 \),

\[
\mathbf{F}(x, y, z) \equiv 2xy\mathbf{i} + (x^2 + 2yz)\mathbf{j} + y^2\mathbf{k}.
\]

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 \( t \).

You might think of \( d\mathbf{R} = \mathbf{r}'(t)\,dt \) to help remember what to do. Then from the definition,

\[
\int_C \mathbf{F} \cdot d\mathbf{R} = \\
\int_0^{2\pi} (2(2t)\sin(t), 4t^2 + 2\sin(t)\cos(t), \sin^2(t)) \cdot (2, -\sin(t), \cos(t))\,dt \\
= \int_0^{2\pi} (8t\sin(t) - (2\sin(t)\cos(t) + 4t^2)\sin(t) + \sin^2(t)\cos(t))\,dt = 16\pi^2 - 16\pi
\]

2. Here is a vector field, \( (y, x^2 + z, 2yz) \) and here is the parametrization of a curve, \( C, \mathbf{R}(t) = (\cos 2t, 2\sin 2t, t) \) where \( t \) goes from 0 to \( \pi/4 \). Find \( \int_C \mathbf{F} \cdot d\mathbf{R} \).

\[
d\mathbf{R} = (-2\sin(2t), 4\cos(2t), 1)\,dt.
\]

Then by the definition,

\[
\int_C \mathbf{F} \cdot d\mathbf{R} = \\
\int_0^{\pi/4} (2\sin(2t), \cos^2(2t) + t, 4t\sin(2t)) \cdot (-2\sin(2t), 4\cos(2t), 1)\,dt \\
= \int_0^{\pi/4} (-4\sin^2(2t) + 4(\cos^2(2t) + t)\cos 2t + 4t\sin 2t)\,dt = \frac{4}{3}
\]
3. Suppose for \( t \in [0, 1] \) the position of an object is given by \( \mathbf{r}(t) = t\mathbf{i} + tj + tk \). Also suppose there is a force field defined on \( \mathbb{R}^3 \),

\[
F(x, y, z) = yz\mathbf{i} + xz\mathbf{j} + xy\mathbf{k}.
\]

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 \( \mathbf{r}(t) = t\mathbf{i} + t^2\mathbf{j} + tk \).

You should get the same answer in this case. This is because the vector field happens to be conservative. (More on this later.)

### 23.2 Conservative Vector Fields, Path Independence

Sometimes the line integral giving the work done by a force field depends only on the endpoints of the curve. This is very nice when it happens because it makes the line integral very easy to compute. It also has great physical significance.

**Definition 23.2.1** A vector field, \( \mathbf{F} \) defined in a three dimensional region is said to be **conservative**\(^1\) if for every piecewise smooth closed curve, \( C \), it follows

\[
\int_C \mathbf{F} \cdot d\mathbf{R} = 0.
\]

**Definition 23.2.2** Let \((x, p_1, \cdots, p_n, y)\) be an ordered list of points in \( \mathbb{R}^p \). Let

\[
p(x, p_1, \cdots, p_n, y)
\]

denote the piecewise smooth curve consisting of a straight line segment from \( x \) to \( p_1 \) and then the straight line segment from \( p_1 \) to \( p_2 \) \( \cdots \) and finally the straight line segment from \( p_n \) to \( y \). 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, \( x, y \in U \), 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 23.2.3** 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 \( x \) to \( y \) in \( U \), then

\[
\int_C \mathbf{F} \cdot d\mathbf{R} = \phi(y) - \phi(x).
\]  

\[ (23.5) \]

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 \( p(x_0, x) \) a polygonal curve from \( x_0 \) to \( x \). Thus the orientation is such that it goes from \( x_0 \) to \( x \). The curve \( p(x, x_0) \) denotes the same set of points but in the opposite order. Suppose first \( \mathbf{F} \) is conservative. Fix \( x_0 \in U \) and let

\[
\phi(x) \equiv \int_{p(x_0, x)} \mathbf{F} \cdot d\mathbf{R}.
\]

\[ ^1 \text{There is no such thing as a liberal vector field.} \]
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

\[
\int_{p(x_0, x)} F \cdot dR + \int_{q(x_0, x)} F \cdot dR = \int_{p(x_0, x)} F \cdot dR - \int_{q(x_0, x)} F \cdot dR
\]

which shows

\[
\int_{p(x_0, x)} F \cdot dR = \int_{q(x_0, x)} F \cdot dR
\]

and that \( \phi \) is well defined. For small \( t \),

\[
\frac{\phi(x + te_i) - \phi(x)}{t} = \frac{\int_{p(x_0, x + te_i)} F \cdot dR - \int_{p(x_0, x)} F \cdot dR}{t}
\]

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 \( x + te_i \) is also contained in \( U \) and so one can take \( p(x, x + te_i)(s) = x + s(te_i) \) for \( s \in [0, 1] \). Therefore, the above difference quotient reduces to

\[
\frac{1}{t} \int_0^1 F(x + s(te_i)) \cdot te_i \, ds = \int_0^1 F_t(x + s(te_i)) \, ds
\]

\[
= F_t(x + s(t)te_i)
\]

by the mean value theorem for integrals. Here \( s_t \) is some number between 0 and 1. By continuity of \( F \), this converges to \( F_t(x) \) as \( t \to 0 \). Therefore, \( \nabla \phi = F \) as claimed.

Conversely, if \( \nabla \phi = F \), then if \( R : [a, b] \to \mathbb{R}^p \) is any \( C^1 \) curve joining \( x \) to \( y \),

\[
\int_a^b F(R(t)) \cdot R'(t) \, dt = \int_a^b \nabla \phi(R(t)) \cdot R'(t) \, dt
\]

\[
= \int_a^b \frac{d}{dt} \phi(R(t)) \, dt
\]

\[
= \phi(R(b)) - \phi(R(a))
\]

\[
= \phi(y) - \phi(x)
\]

and this verifies 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 \( \phi(x) \) 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.

23.2.1 Finding The Scalar Potential, (Recover The Function From Its Gradient)

Example 23.2.4 Let \( F(x, y, z) = (\cos x - yz \sin (xz), \cos (xz), -yx \sin (xz)) \). Let \( C \) be a piecewise smooth curve which goes from \((\pi, 1, 1)\) to \((\pi, 3, 2)\). Find \( \int_C F \cdot dR \).
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

\[
\phi_x = \cos x - y (\sin x) z
\]

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 23.2.5** Let \( \mathbf{F} (x, y, z) = (x, y^2 x, z) \). 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.

**Example 23.2.6** Let \( \mathbf{F} (x, y, z) = (2yx + 1 + y, x^2 + x, 1) \). Find a scalar potential for \( \mathbf{F} \) if it exists.

You need \( \phi_x = 2yx + 1 + y \) and so \( \phi = yx^2 + x + yx + \psi (y, z) \). Then you need \( \phi_y = x^2 + x + yx + \psi (y, z) \) which shows \( \psi_y = 0 \) and so \( \psi = \psi (z) \). Hence \( \phi = yx^2 + x + yx + \psi (z) \) Now finally, \( \phi_z = \psi' (z) = 1 \) and so \( \psi (z) = z \) will work. A scalar potential is \( \phi (x, y, z) = yx^2 + x + yx + z \).

**Example 23.2.7** Let \( \mathbf{F} (x, y, z) = (1, 2yz + z \cos y, y^2 + \sin y) \). Find a scalar potential for \( \mathbf{F} \) if it exists.

You need \( \phi_x = 1 \) and so \( \phi = x + \psi (y, z) \). Then you need \( \phi_y = \psi_y = 2yz + z \cos y \) and so \( \psi = y^2 z + z \sin y + g (z) \). Hence \( \phi = x + y^2 z + z \sin y + g (z) \) and you still don’t know \( g \). But you must have \( \phi_z = y^2 + \sin y + g' (z) = z^2 + \sin y \) and so \( g \) is a constant. You can take it to equal zero. Hence \( \phi = x + y^2 z + z \sin y \) is a scalar potential.

When you are finding one of these scalar potentials, be sure to check your work. Take what you think is the answer and find its gradient. If you get the given vector field, rejoice. If not, it is wrong. Start over again.
Example 23.2.8 Let the vector field, \( \mathbf{F} \) be given in Example 23.2.7. Find \( \int_C \mathbf{F} \cdot d\mathbf{R} \) where \( C \) is an oriented curve which goes from \((0, \pi, 2)\) to \((1, \pi/2, 2)\).

This is very easy. It is just \( \phi(1, \pi/2, 2) - \phi(0, \pi, 2) \) where \( \phi \) is the scalar potential in this example. Thus it equals

\[
\left( 1 + \left( \frac{\pi}{2} \right)^2 + 2 \right) - (\pi^2 + 2) = 3 - \frac{3}{2} \pi^2
\]

23.2.2 Terminology

For a vector field, \( \mathbf{F}(x, y, z) = F_1(x, y, z) \mathbf{i} + F_2(x, y, z) \mathbf{j} + F_3(x, y, z) \mathbf{k} \), \( \mathbf{F} \) is called conservative if it is the gradient of a scalar potential. Thus \( \mathbf{F} \) is conservative if there exists a scalar function, \( \phi \) such that \( \nabla \phi = \mathbf{F} \). This was discussed above. Another way to say this is that the differential form \( F_1 dx + F_2 dy + F_3 dz \) is exact. This terminology holds with obvious modifications in any number of dimensions.

23.3 Divergence, Gradient, Curl

23.3.1 Divergence

There are several things you can do with vector fields and scalar fields. Earlier the gradient was discussed. The gradient starts with a scalar field and produces a vector field. The divergence starts with a vector field and produces a scalar field. The curl starts with a vector field in \( \mathbb{R}^3 \) and produces another vector field in \( \mathbb{R}^3 \).

Definition 23.3.1 Let \( \mathbf{f} : U \to \mathbb{R}^p \) for \( U \subseteq \mathbb{R}^p \) denote a vector field. A scalar valued function is called a scalar field. The function, \( \mathbf{f} \) is called a \( \mathcal{C}^k \) vector field if the function, \( \mathbf{f} \) is a \( \mathcal{C}^k \) function. For a \( \mathcal{C}^1 \) vector field, as just described \( \nabla \cdot \mathbf{f}(x) \equiv \text{div} \mathbf{f}(x) \) known as the divergence, is defined as

\[
\nabla \cdot \mathbf{f}(x) \equiv \text{div} \mathbf{f}(x) \equiv \sum_{i=1}^{p} \partial f_i(x) \partial x_i(x).
\]

Using the repeated index 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 \( \mathbf{f} \). Also

\[
\nabla^2 \mathbf{f} \equiv \nabla \cdot (\nabla \mathbf{f}).
\]

This last symbol is important enough that it is given a name, the Laplacian. It is also denoted by \( \Delta \). Thus \( \nabla^2 \mathbf{f} = \Delta \mathbf{f} \). In addition for \( \mathbf{f} \) a vector field, the symbol \( \mathbf{f} \cdot \nabla \) is defined as a “differential operator” in the following way.

\[
\mathbf{f} \cdot \nabla (g) \equiv f_1(x) \frac{\partial g(x)}{\partial x_1} + f_2(x) \frac{\partial g(x)}{\partial x_2} + \cdots + f_p(x) \frac{\partial g(x)}{\partial x_p}.
\]

Thus \( \mathbf{f} \cdot \nabla \) takes vector fields and makes them into new vector fields.
CHAPTER 23. LINE INTEGRALS

This definition is in terms of a given coordinate system but later a coordinate free definition of div is presented. For now, everything is defined in terms of a given Cartesian coordinate system. The divergence has profound physical significance and this will be discussed later. For now it is important to understand how to find it. 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.

**Example 23.3.2** Let \( f(x) = x y i + (z - y) j + (\sin(x) + z) k \). Find \( \text{div} \ f \)

First the divergence of \( f \) is

\[
\frac{\partial (x y)}{\partial x} + \frac{\partial (z - y)}{\partial y} + \frac{\partial (\sin(x) + z)}{\partial z} = y + (-1) + 1 = y.
\]

### 23.3.2 Gradient

For a scalar field, \( f \), \( \nabla f \) is a vector field described earlier.

**Definition 23.3.3** Suppose \( f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R} \) 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

\[
\nabla f(x) = (f_{x_1}(x), f_{x_2}(x), \ldots, f_{x_n}(x))^T.
\]

**Example 23.3.4** Find \( \nabla (x^2 y + \sin(z)) \).

This is as just defined \((2xy, x^2, \cos z)\)

### 23.3.3 Curl

Here the important concepts of curl is defined.

**Definition 23.3.5** Let \( f : U \rightarrow \mathbb{R}^3 \) for \( U \subseteq \mathbb{R}^3 \) denote a vector field. The curl of the vector field yields another vector field and it is defined as follows.

\[
(\text{curl}(f)(x))_i = \langle \nabla \times f(x) \rangle_i = \varepsilon_{ijk} \partial_j f_k(x)
\]

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 \( i^{\text{th}} \) Cartesian component of the vector, \( \text{curl}(f)(x) \). If you didn’t read the notation involving repeated indices and the permutation symbol, use the following. The curl is evaluated by expanding the following determinant along the top row.

\[
\begin{vmatrix}
  i & j & k \\
  \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\
  f_1(x,y,z) & f_2(x,y,z) & f_3(x,y,z)
\end{vmatrix}.
\]

Note the similarity with the cross product. Sometimes the curl is called rot. (Short for rotation not decay.)

This definition is in terms of a given coordinate system but later coordinate free definitions of the curl is presented. For now, everything is defined in terms of a given Cartesian coordinate system. The curl has profound physical significance and this will be discussed later. For now it is important to understand how to find it. Be sure you understand that for \( f \) a vector field, \( \text{curl} \ f \) is another vector field.
Example 23.3.6 Let \( f(x) = xy + (z - y)j + (\sin(x) + z)k \). Find \( \text{curl} f \).

\( \text{curl} f \) is obtained by evaluating

\[
\begin{vmatrix}
i & j & k \\
\frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\
x y & z - y & \sin(x) + z
\end{vmatrix}
\]

\[
i \left( \frac{\partial}{\partial y} (\sin(x) + z) - \frac{\partial}{\partial z} (z - y) \right) -
\]

\[
j \left( \frac{\partial}{\partial x} (\sin(x) + z) - \frac{\partial}{\partial z} (xy) \right) +
\]

\[
k \left( \frac{\partial}{\partial x} (z - y) - \frac{\partial}{\partial y} (xy) \right) = -i - \cos(x)j - xk.
\]

23.3.4 Identities Involving Curl, Divergence, and Gradient

There are a number of identities involving these operations and they are listed in the following theorem.

Theorem 23.3.7 Let \( u \) be \( C^2 \) and let \( f, g \) also be \( C^2 \). Then the following identities hold.

1. \( \nabla \cdot (\nabla u) = \nabla^2 u \)
2. \( \nabla \cdot (\nabla \times f) = 0 \)
3. \( \nabla \times (\nabla u) = 0 \)
4. \( \nabla \times (\nabla \times f) = \nabla (\nabla \cdot f) - \nabla^2 f \)
5. \( \nabla \cdot (uf) = u \nabla \cdot f + \nabla (u \cdot f) \)
6. \( \nabla \cdot (f \times g) = g \cdot (\nabla \times f) - f \cdot (\nabla \times g) \)
7. \( \nabla \times (f \times g) = (\nabla \cdot g) f - (\nabla \cdot f) g + (g \cdot \nabla) f - (f \cdot \nabla) g \)

Proof: The first one is really the definition of \( \nabla^2 \). The second follows from a simple computation as does the third. Just plug in to the formulas and use equality of mixed partial derivatives and it will all fall out. The fourth one is much harder but is quite easy if you use the repeated index summation convention along with the reduction identities for the permutation symbol.

\[
(\nabla \times (\nabla \times f))_i = \varepsilon_{ijk} \partial_j (\nabla \times f)_k \\
= \varepsilon_{ijk} \partial_j \varepsilon_{krs} \partial_r F_s = \varepsilon_{kij} \varepsilon_{krs} \partial_j \partial_r F_s \\
= (\delta_{jr} \delta_{ks} - \delta_{js} \delta_{kr}) \partial_j \partial_r F_s = \partial_j \partial_i F_j - \partial_j \partial_j F_i \\
= \partial_i (\nabla \cdot F) - \nabla^2 F_i
\]

Hence

\[
\nabla \times (\nabla \times f) = \nabla (\nabla \cdot F) - \nabla^2 f
\]

The fifth claim is easy from the definition. Claim 6 is also fairly easy from the permutation symbol identities.

\[
\nabla \cdot (f \times g) = \partial_i (\varepsilon_{ijk} f_j g_k) = \varepsilon_{ijk} ((\partial_i f_j) g_k + f_j (\partial_i g_k)) \\
= \varepsilon_{kij} (\partial_i f_j) g_k - f_j \varepsilon_{ijk} (\partial_i g_k) \\
= g \cdot \nabla \times f - f \cdot \nabla \times g
\]
Consider now claim 7.

\[
\nabla \times (f \times g)_i = \varepsilon_{ijk} \partial_j (f \times g)_k
\]

\[
= \varepsilon_{ijk} \partial_j \varepsilon_{krs} f_r g_s
\]

\[
= \varepsilon_{kij} \varepsilon_{krs} \partial_j (f_r g_s)
\]

\[
= (\delta_{ir} \delta_{js} - \delta_{is} \delta_{jr}) \partial_j (f_r g_s)
\]

\[
= \partial_j (f_i g_j) - \partial_j (f_j g_i)
\]

\[
= (\partial_j g_j) f_i + g_j \partial_j f_i - (\partial_j f_j) g_i - f_j (\partial_j g_i)
\]

\[
= (\nabla \cdot g) f + (g \cdot \nabla) (f) - (\nabla \cdot f) g - (f \cdot \nabla) (g)_i
\]

Recall \((g \cdot \nabla) f \equiv \sum_j g_j (\partial_j f)\). This proves the identities.

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 59 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.

### 23.3.5 Vector Potentials

One of the above identities says \(\nabla \cdot (\nabla \times f) = 0\). Suppose now \(\nabla \cdot g = 0\). Does it follow that there exists \(f\) such that \(g = \nabla \times f\)? It turns out that this is usually the case and when such an \(f\) exists, it is called a vector potential. Here is one way to do it, assuming everything is defined so the following formulas make sense.

\[
f(x, y, z) = \left( \int_0^z g_2(x, y, t) \, dt, -\int_0^z g_1(x, y, t) \, dt + \int_0^x g_3(t, y, 0) \, dt, 0 \right)^T.
\]

(23.6)

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.

\[
\frac{\partial}{\partial x} \int_a^b h(x, t) \, dt = \int_a^b \frac{\partial h}{\partial x}(x, t) \, dt.
\]

(23.7)

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 23.3.8** Suppose \(h\) and \(\frac{\partial h}{\partial y}\) are continuous on the rectangle \(R = [c, d] \times [a, b]\). Then \(23.8\) holds.
Part VIII

Green’s Theorem, Integrals On Surfaces
Chapter 24

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 24.0.9** Let \( F(x, y) \equiv (P(x, y), Q(x, y)) \) 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

\[
\int_{\partial U} P \, dx + Q \, dy = \int_{\partial U} \mathbf{F} \cdot d\mathbf{R} = \int_U \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) \, dA. \tag{24.1}
\]

**Proof:** First consider the right side of (24.1)

\[
\int_U \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) \, dA
\]
\[ = \int_c^d \int_{l(y)}^{r(y)} \frac{\partial Q}{\partial x} \, dx \, dy - \int_a^b \int_{b(x)}^{r(x)} \frac{\partial P}{\partial y} \, dy \, dx \]
\[ = \int_c^d (Q(r(y), y) - Q(l(y), y)) \, dy + \int_a^b (P(x, b(x)) - P(x, t(x))) \, dx. \quad (24.2) \]

Now consider the left side of \(24.1\). Denote by \(V\) the vertical parts of \(\partial U\) and by \(H\) the horizontal parts.

\[
\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} = 
\]
\[
= \int_c^d (0, Q(r(s), s)) \cdot (r'(s), 1) \, ds + \int_H (0, Q(r(s), s)) \cdot (\pm 1, 0) \, ds \\
- \int_c^d (0, Q(l(s), s)) \cdot (l'(s), 1) \, ds + \int_a^b (P(s, b(s)), 0) \cdot (1, b'(s)) \, ds \\
+ \int_V (P(s, b(s)), 0) \cdot (0, \pm 1) \, ds - \int_a^b (P(s, t(s)), 0) \cdot (1, t'(s)) \, ds \\
= \int_c^d Q(r(s), s) \, ds - \int_c^d Q(l(s), s) \, ds + \int_a^b P(s, b(s)) \, ds - \int_a^b P(s, t(s)) \, ds
\]

which coincides with (24.2). This proves the lemma.

**Corollary 24.0.10** Let everything be the same as in Lemma 24.0.9 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 \(U \equiv \bigcup_{i=1}^m U_i\), the union of these sets. This is because

\[
\int_U \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) \, dA = 
\]
\[
= \sum_{k=1}^m \int_{U_k} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) \, dA \\
= \sum_{k=1}^m \int_{\partial U_k} \mathbf{F} \cdot d\mathbf{R} = \int_{\partial U} \mathbf{F} \cdot d\mathbf{R}
\]

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.

\[
\int_{\partial V} \mathbf{F} \cdot d\mathbf{R} = \int_{V} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA
\]

\[
= \int_{U} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA + \int_{V \setminus U} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA
\]

and so

\[
\int_{V \setminus U} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \int_{\partial V} \mathbf{F} \cdot d\mathbf{R} - \int_{\partial U} \mathbf{F} \cdot d\mathbf{R}
\]

which equals

\[
\int_{\partial (V \setminus U)} \mathbf{F} \cdot d\mathbf{R}
\]

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 24.0.11 (Green’s Theorem)** Let \( U \) be an open set in the plane and let \( \partial U \) be piecewise smooth and let \( \mathbf{F}(x, y) = (P(x, y), Q(x, y)) \) be a \( C^1 \) vector field defined near \( U \). Then it is often the case that

\[
\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} = \int_{U} \left( \frac{\partial Q}{\partial x} (x, y) - \frac{\partial P}{\partial y} (x, y) \right) dA.
\]

\(^1\)For a general version see the advanced calculus book by Apostol. The general versions involve the concept of a rectifiable Jordan curve. You need to be able to take the area integral and to take the line integral around the boundary.
24.1 An Alternative Explanation Of Green’s Theorem

Consider the following picture.

In this picture \( \mathbf{n} \) is the unit outer normal to \( U \) and the vector, \( \mathbf{T} \) shown in the picture is the unit tangent vector in the direction of counter clockwise motion around \( U \). To see that it really does point in the correct direction, take the cross product, \( (n_x, n_y, 0) \times (-n_y, n_x, 0) \). This equals \( \mathbf{k} \). Applying the right hand rule, this shows the vector, \( (-n_y, n_x) \) really does point in the direction indicated by the picture.

Next I will establish Gauss’ theorem for regions like \( U \). The boundary of \( U \) is denoted by \( \partial U \).

**Lemma 24.1.1 (Gauss)** Let \((H(x,y), K(x,y))\) be a \( C^1 \) vector field defined near \( U \). Then for \( \mathbf{n} \) the unit outer normal,

\[
\int_U (H_x + K_y) \, dA = \int_{\partial U} (H, K) \cdot d\mathbf{l}.
\]

**Proof:** A parametrization for the top is \((x, T(x))\) and a parametrization for the bottom is \((x, B(x))\) where in both cases, \( x \in [a, b] \). Thus \( dl = \sqrt{1 + T'(x)^2} \, dx \) on the top and \( dl = \sqrt{1 + B'(x)^2} \, dx \) on the bottom. Thus also, on the top, you can find the exterior normal by considering it as the level surface, \( y - T(x) = 0 \). Thus a unit normal to this surface is

\[
\mathbf{n} = \frac{(-T'(x), 1)}{\sqrt{1 + T'(x)^2}} = (n_x, n_y)
\]

and you see that since the \( y \) component is positive, it is the outer normal, pointing away from \( U \). Similarly, the unit outer normal on the bottom is given by

\[
\mathbf{n} = \frac{(B'(x), -1)}{\sqrt{1 + B'(x)^2}} = (n_x, n_y)
\]
First consider
\[ \int_U K_y dA = \int_a^b \int_{T(x)}^{B(x)} K_y dydx = \int_a^b (K(x, T(x)) - K(x, B(x))) dx \]
\[ = \int_a^b K(x, T(x)) dx - \int_a^b K(x, B(x)) dx \]
\[ = \int_a^b K(x, T(x)) \frac{1}{\sqrt{1 + T'(x)^2}} dx \]
\[ + \int_a^b K(x, B(x)) \frac{-1}{\sqrt{1 + B'(x)^2}} dx \]
\[ = \int_{\partial U} Kn_y dl \]

Similar reasoning shows that
\[ \int_U H_x dA = \int_{\partial U} H n_x dl \]

Therefore, this proves the Lemma because from the above,
\[ \int_U (H_x + K_y) dA = \int_U H_x dA + \int_U K_y dA = \int_{\partial U} H n_x dl + \int_{\partial U} K n_y dl \]
\[ = \int_{\partial U} (H, K) \cdot n dl \]

Now this theorem holds for many regions much more general than the one shown. In fact, it holds for any region which is made up by pasting together regions like the above. This is because the area integrals add and the integrals on the parts of the boundary which are shared by two pieces cancel due to the fact they have the exterior normals which are in opposite directions. 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.

\[ \begin{array}{c}
V_1 \\
V_2
\end{array} \]

This theorem is also called the divergence theorem. This is because the divergence of the vector field, \((H(x, y), K(x, y))\) is defined as \(H_x(x, y) + K_y(x, y)\).

**Theorem 24.1.2 (Green’s Theorem)** Let \( U \) be any bounded open set in \( \mathbb{R}^2 \) for which the above Gauss’ theorem holds and let
\[ \mathbf{F}(x, y) \equiv (P(x, y), Q(x, y)) \]
be a \( C^1 \) vector field defined near \( U \). Then
\[ \int_U (Q_x - P_y) dA = \int_{\partial U} \mathbf{F} \cdot d\mathbf{R} \]
CHAPTER 24. GREEN’S THEOREM

where the line integral is oriented in the counter clockwise direction.

Proof: If \( r(t) \) is a parametrization of \( \partial U \) near a point on \( \partial U \), then recall the unit tangent vector, \( T \) as shown in the above picture satisfies \( |r'(t)|T = r'(t) \). Thus \( F \cdot dR \) is of the form \( F(r(t)) \cdot r'(t) \, dt = F \cdot T |r'(t)| \, dt = F \cdot T \, dl \) because \( dl = |r'(t)| \, dt \). Then using Lemma 24.1.1 and letting \((H,K) = (Q,-P)\)

\[
\int_{U} (Q_x - P_y) \, dA = \int_{U} (H_x + K_y) \, dA \\
= \int_{\partial U} (H,K) \cdot (n_x,n_y) \, dl \\
= \int_{\partial U} (Q,-P) \cdot (n_x,n_y) \, dl \\
= \int_{\partial U} (P,Q) \cdot (-n_y,n_x) \, dl \\
= \int_{\partial U} F \cdot T \, dl = \int_{\partial U} F \cdot dR.
\]

This proves Green’s theorem.

24.2 Area And Green’s Theorem

Proposition 24.2.1 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 \( F(x,y) = \frac{1}{2} (-y,x), (0,x) \), or \((y,0)\).

Proof: This follows immediately from Green’s theorem.

Example 24.2.2 Use Proposition 24.2.1 to find the area of the ellipse

\[
\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1.
\]

You can parametrize the boundary of this ellipse as

\[
x = a \cos t, \quad y = b \sin t, \quad t \in [0,2\pi].
\]

Then from Proposition 24.2.1

\[
\text{Area equals} = \frac{1}{2} \int_{0}^{2\pi} (-b \sin t, a \cos t) \cdot (-a \sin t, b \cos t) \, dt \\
= \frac{1}{2} \int_{0}^{2\pi} (ab) \, dt = \pi ab.
\]

Example 24.2.3 Find \( \int_{\partial U} F \cdot dR \) where \( U \) is the set, \( \{(x,y) : x^2 + 3y^2 \leq 9\} \) and \( F(x,y) = (y,-x) \).

One way to do this is to parametrize the boundary of \( U \) and then compute the line integral directly. It is easier to use Green’s theorem. The desired line integral equals

\[
\int_{U} ((-1) - 1) \, dA = -2 \int_{U} dA.
\]

Now \( U \) is an ellipse having area equal to \( 3\sqrt{3} \) and so the answer is \(-6\sqrt{3}\).
Example 24.2.4 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 24.2.5 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 4\} and $\mathbf{F}(x, y) = (x \sin y, y \sin x)$.

From Green’s theorem this line integral equals
\[
\int_2^4 \int_x^4 (y \cos x - x \cos y) \, dy \, dx = 4 \cos 2 - 8 \cos 4 - 8 \sin 2 - 4 \sin 4
\]
Chapter 25

The Integral On Two Dimensional Surfaces In $\mathbb{R}^3$

25.1 Parametrically Defined Surfaces

Definition 25.1.1 Let $S$ be a subset of $\mathbb{R}^3$. Then $S$ is a smooth surface if there exists an open set, $U \subseteq \mathbb{R}^2$ and a $C^1$ function, $r$ defined on $U$ such that $r(U) = S$, $r$ is one to one, and for all $(u,v) \in U$,

$$r_u \times r_v \neq 0. \tag{25.1}$$

This last condition ensures that there is always a well defined normal on $S$. This function, $r$ is called a parametrization of the surface. It is just like a parametrization of a curve but here there are two parameters, $u,v$.

One way to think of this is that there is a piece of rubber occupying $U$ in the plane and then it is taken and stretched in three dimensions. This gives $S$. Here is an interesting example which is already familiar.

Example 25.1.2 Let $(\phi, \theta) \in (0, \pi) \times (0, 2\pi)$ and for such $(\phi, \theta)$,

$$r(\phi, \theta) = \begin{pmatrix} 2 \sin(\phi) \cos(\theta) \\ 2 \sin(\phi) \sin(\theta) \\ 2 \cos(\phi) \end{pmatrix}$$

This gives most of a sphere of radius 2 for $S$. You should check condition (25.1). You will find that $|r_\phi \times r_\theta| = 4 \sin(\phi) \neq 0$.

Example 25.1.3 Let $R > r$. Consider

$$r(u,v) = \begin{pmatrix} (R + r \cos(u)) \cos(v) \\ (R + r \cos(u)) \sin(v) \\ r \sin(u) \end{pmatrix}$$

where $(u,v) \in (0, 2\pi) \times (0, 2\pi)$. This surface is most of the surface of a torus (donut) with small radius equal to $r$. It is obtained by revolving the circle of radius $r$ centered at $(R,0,0)$ about the $z$ axis. Here is a picture.
In the above I have assumed $U$ is open. However, this is generalized later. It is amazing how far this can be generalized in applications to integration.

In general, if you fix $u$ and consider $r(u, v)$ as a function of $v$, this yields a smooth curve which lies in the surface, $S$. By fixing different values of $u$ you obtain many different curves in $S$. Similarly you can fix $v$ and consider $r(u, v)$ as a function of $u$. The curves which result in this way are called a net for the surface. This is the way a computer graphs a surface. It graphs lots of different curves as just described. You can see this in the above picture of a torus. The curves which make up the shape shown correspond to one of the variables in the parametrization being fixed. Now at a point, $r(u, v)$ of $S$, there are two vectors tangent to $S$ at this point, $r_u(u, v)$ and $r_v(u, v)$. These two vectors determine a plane which can be considered tangent to the surface at the point, $r(u, v)$. You can find an equation for this plane if you can obtain a normal vector. However, this is easy. You simply take $r_v \times r_u$ to obtain a vector which is normal to the tangent plane. Here is a picture. The two curves correspond to $u \to r(u, v)$ and $v \to r(u, v)$. The vectors $r_u$ and $r_v$ are tangent to the respective curves as shown. Then taking the cross product gives a normal to the surface at that point.

Example 25.1.4 Let $S$ be the surface defined in Example 25.1.3 in which $R = 2$ and $r = 1$. Find a tangent plane to the point $r \left( \frac{\pi}{4}, \frac{\pi}{4} \right)$.

This point is $\left( \sqrt{2} + \frac{1}{2}, \sqrt{2} + \frac{1}{2}, \frac{\sqrt{2}}{2} \right)$. I only need to find a normal vector in order to find
the plane.

\[
\mathbf{r}_u \times \mathbf{r}_v = \begin{pmatrix}
-\sin u \cos v \\
-\sin u \sin v \\
\cos u
\end{pmatrix} \times \begin{pmatrix}
-(2 + \cos u) \sin v \\
(2 + \cos u) \cos v \\
0
\end{pmatrix} = \begin{pmatrix}
-\cos u (2 + \cos u) \cos v \\
-\cos u (2 + \cos u) \sin v \\
-(\sin u \cos^2 v) (2 + \cos u) - (\sin u \sin^2 v) (2 + \cos u)
\end{pmatrix}
\]

Now plugging in the desired values of \( u \) and \( v \), a normal vector is

\[
\begin{pmatrix}
-1 - \frac{1}{4} \sqrt{2} \\
-1 - \frac{1}{4} \sqrt{2} \\
-\sqrt{2} - \frac{1}{2}
\end{pmatrix}.
\]

I don’t like the minus signs so the normal vector I will use is

\[
\left(1 + \frac{1}{4} \sqrt{2}, 1 + \frac{1}{4} \sqrt{2}, \sqrt{2} + \frac{1}{2}\right)^T.
\]

Now it follows the equation of the tangent plane is

\[
\left(1 + \frac{1}{4} \sqrt{2}\right) (x - \sqrt{2} - \frac{1}{2}) + \left(1 + \frac{1}{4} \sqrt{2}\right) (y - \sqrt{2} - \frac{1}{2}) + \left(\sqrt{2} + \frac{1}{2}\right) (z - \frac{1}{2} \sqrt{2}) = 0.
\]

You could simplify this if you wanted.

\[
\left(1 + \frac{1}{4} \sqrt{2}\right) x + \left(1 + \frac{1}{4} \sqrt{2}\right) y + \left(\sqrt{2} + \frac{1}{2}\right) z = \frac{5}{2} \sqrt{2} + 3.
\]

### 25.2 The Two Dimensional Area In \( \mathbb{R}^3 \)

Consider the boundary of some three dimensional region such that a function, 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. It all depends on the following fundamental definition which is just a review of the fact presented earlier that the area of a parallelogram determined by two vectors in \( \mathbb{R}^3 \) is the norm of the cross product of the two vectors.

**Definition 25.2.1** Let \( \mathbf{u}_1, \mathbf{u}_2 \) be vectors in \( \mathbb{R}^3 \). The 2 dimensional parallelogram determined by these vectors will be denoted by \( P(\mathbf{u}_1, \mathbf{u}_2) \) and it is defined as

\[
P(\mathbf{u}_1, \mathbf{u}_2) \equiv \left\{ \sum_{j=1}^{2} s_j \mathbf{u}_j : s_j \in [0, 1] \right\}.
\]

Then the area of this parallelogram is

\[
\text{area } P(\mathbf{u}_1, \mathbf{u}_2) \equiv |\mathbf{u}_1 \times \mathbf{u}_2|.
\]
Suppose then that $x = r(u)$ where $u \in U$, a subset of $\mathbb{R}^2$ and $x$ is a point in $V$, a subset of 3 dimensional space. Thus, letting the Cartesian coordinates of $x$ be given by $x = (x_1, x_2, x_3)^T$, each $x_i$ being a function of $u$, an infinitesimal rectangle located at $u_0$ corresponds to an infinitesimal parallelogram located at $r(u_0)$ which is determined by the 2 vectors $\{\frac{\partial r(u_0)}{\partial u_i} du_i\}_{i=1}^2$, each of which is tangent to the surface defined by $x = r(u)$. (No sum on the repeated index.)

From Definition 25.2.1, the volume of this infinitesimal parallelepiped located at $r(u_0)$ is given by

$$dS \equiv \left| \frac{\partial r(u_0)}{\partial u_1} \times \frac{\partial r(u_0)}{\partial u_2} \right| du_1 du_2 \quad (25.2)$$

$$= \left| r_{u_1} \times r_{u_2} \right| du_1 du_2 \quad (25.3)$$

It might help to think of a lizard. The infinitesimal parallelogram is 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 If the scales are small enough, their sum would serve as a good approximation to the area of the lizard.\[\text{This beautiful lizard is a } Sceloporus\ magister. \text{ It was photographed by C. Riley Nelson who is in the Zoology department at Brigham Young University © 2004 in Kane Co. Utah. The lizard is a little less than one foot in length.}\]
This motivates the following fundamental procedure which I hope is extremely familiar from the earlier material.

**Procedure 25.2.2** Suppose $U$ is a subset of $\mathbb{R}^2$ and suppose $r : U \rightarrow r(U) \subseteq \mathbb{R}^3$ is a one to one and $C^1$ function. Then if $h : r(U) \rightarrow \mathbb{R}$, define the 2 dimensional surface integral, $\int_{r(U)} h(x) \, dA$ according to the following formula.

$$\int_{r(U)} h(x) \, dS = \int_U h(r(u)) |r_{u_1}(u) \times r_{u_2}(u)| \, du_1 du_2.$$ 

**Definition 25.2.3** It is customary to write $|r_{u_1}(u) \times r_{u_2}(u)| = \frac{\partial (x_1,x_2,x_3)}{\partial (u_1,u_2)}$ because this new notation generalizes to far more general situations for which the cross product is not defined. For example, one can consider three dimensional surfaces in $\mathbb{R}^8$.

First here is a simple example where the surface is actually in the plane.

**Example 25.2.4** Find the area of the region labelled $A$ in the following picture. The two circles are of radius 1, one has center $(0,0)$ and the other has center $(1,0)$.

![Diagram of two circles](image.png)

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 $\theta \in [-\frac{\pi}{3},\frac{\pi}{3}]$ 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,

$$\int_{U} 1 \, dV = \int_{-\pi/3}^{\pi/3} \int_{1}^{2 \cos \theta} \frac{\partial (x_1,x_2,x_3)}{\partial (\theta,r)} \, dr \, d\theta.$$ 

It is necessary to find $\frac{\partial (x_1,x_2)}{\partial (\theta,r)}$. The mapping $r : U \rightarrow \mathbb{R}^2$ takes the form

$$r (\theta, r) = (r \cos \theta, r \sin \theta)^T.$$ 

Here $x_3 = 0$ and so

$$\frac{\partial (x_1,x_2,x_3)}{\partial (\theta,r)} = \begin{vmatrix} i & j & k \\ \frac{\partial x_1}{\partial \theta} & \frac{\partial x_2}{\partial \theta} & \frac{\partial x_3}{\partial \theta} \\ \frac{\partial x_1}{\partial r} & \frac{\partial x_2}{\partial r} & \frac{\partial x_3}{\partial r} \end{vmatrix} = \begin{vmatrix} i & j & k \\ -r \sin \theta & r \cos \theta & 0 \\ \cos \theta & \sin \theta & 0 \end{vmatrix} = r$$ 

Therefore, the area element is $r \, dr \, d\theta$. It follows the desired area is

$$\int_{-\pi/3}^{\pi/3} \int_{1}^{2 \cos \theta} r \, dr \, d\theta = \frac{1}{2} \sqrt{3} + \frac{1}{3} \pi.$$ 

Notice how the area element reduced to the area element for polar coordinates.
Example 25.2.5 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 \( x = r(u) \). This is easy to do if you let \( u = (x, y) \). Then \( r(x, y) = (x, y, x^2) \). If you like, let \( x = u_1 \) and \( y = u_2 \).

What is \( \frac{\partial(x_1, x_2, x_3)}{\partial(x, y)} = |r_x \times r_y| \)?

\[
\begin{align*}
    r_x &= \begin{pmatrix} 1 \\ 0 \\ 2x \end{pmatrix}, \\
    r_y &= \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}
\end{align*}
\]

and so

\[
|r_x \times r_y| = \begin{vmatrix} 1 & 0 & 2x \\ 0 & 1 & 0 \\ 2x & 0 & 0 \end{vmatrix} = \sqrt{1 + 4x^2}
\]

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

\[
\int_{r(U)} dA = \int_0^1 \int_0^1 \sqrt{1 + 4x^2} \, dx \, dy = \frac{1}{2} \sqrt{5} - \frac{1}{4} \ln \left( -2 + \sqrt{5} \right)
\]

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, \( x = r(u) \) 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

\[
\sin \left( x^2 + \ln (7 + y^2 \sin x) \right) + \sin (zx) \, e^z = 11 \sin (xyz)\?
\]

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 parametrization at least locally. The problem is, 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. This is a good example which illustrates the gulf between theory and practice.

Example 25.2.6 Let \( U = [0, 12] \times [0, 2\pi] \) and let \( r : U \to \mathbb{R}^3 \) be given by \( r(t, s) = (2 \cos t + \cos s, 2 \sin t + \sin s, t)^T \). Find a double integral for the surface area. A graph of this surface is drawn below.
25.2. THE TWO DIMENSIONAL AREA IN $\mathbb{R}^3$

It looks like something you would use to make sausages\(^2\). Anyway,

$$
\mathbf{r}_t = \begin{pmatrix} -2 \sin t \\ 2 \cos t \\ 1 \end{pmatrix}, \quad \mathbf{r}_s = \begin{pmatrix} -\sin s \\ \cos s \\ 0 \end{pmatrix}
$$

and

$$
\mathbf{r}_t \times \mathbf{r}_s = \begin{pmatrix} -\cos s \\ -\sin s \\ -2 \sin t \cos s + 2 \cos t \sin s \end{pmatrix}
$$

and so

$$
\frac{\partial (x_1, x_2, x_3)}{\partial (t, s)} = |\mathbf{r}_t \times \mathbf{r}_s| = \sqrt{5 - 4 \sin^2 t \sin^2 s - 8 \sin t \sin s \cos t \cos s - 4 \cos^2 t \cos^2 s}.
$$

Therefore, the desired integral giving the area is

$$
\int_0^{2\pi} \int_0^{12} \sqrt{5 - 4 \sin^2 t \sin^2 s - 8 \sin t \sin s \cos t \cos s - 4 \cos^2 t \cos^2 s} \, dt \, ds.
$$

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. It is important in some physical applications to get the number even when you can’t find the antiderivative. Here is an example for which you will be able to find the integrals.

**Example 25.2.7** Let $U = [0, 2\pi] \times [0, 2\pi]$ and for $(t, s) \in U$, let

$$
\mathbf{r}(t, s) = (2 \cos t + \cos t \cos s, -2 \sin t - \sin t \cos s, \sin s)^T.
$$

Find the area of $\mathbf{r}(U)$. This is the surface of a donut shown below. The fancy name for this shape is a torus.

\(^2\)At Wolverth’s in Hancock Michigan, they make excellent sausages and hot dogs. The best are made from “natural casings” which are the linings of intestines.
To find its area,
\[ \mathbf{r}_t = \begin{pmatrix} -2 \sin t - \sin t \cos s \\ -2 \cos t - \cos t \cos s \\ 0 \end{pmatrix}, \quad \mathbf{r}_s = \begin{pmatrix} -\cos t \sin s \\ \sin t \sin s \\ \cos s \end{pmatrix}, \]
and so \(|\mathbf{r}_t \times \mathbf{r}_s| = (\cos s + 2)\) so the area element is \((\cos s + 2) \, ds \, dt\) and the area is
\[ \int_0^{2\pi} \int_0^{2\pi} (\cos s + 2) \, ds \, dt = 8\pi^2. \]

**Example 25.2.8** Let \(U = [0, 2\pi] \times [0, 2\pi]\) and for \((t, s) \in U\), let
\[ \mathbf{r}(t, s) = (2 \cos t + \cos t \cos s, -2 \sin t - \sin t \cos s, \sin s)^T. \]
Find
\[ \int_{\mathbf{r}(U)} h \, dV \]
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
\[ \int_{\mathbf{r}(U)} h \, dA = \int_0^{2\pi} \int_0^{2\pi} \left( \frac{x \text{ on the surface}}{2 \cos t + \cos t \cos s} \right)^2 (\cos s + 2) \, ds \, dt = 22\pi^2. \]

**Example 25.2.9** Let \(U = [-5, 5] \times [0, 3\pi]\) and for \((s, t) \in U\), let
\[ \mathbf{r}(s, t) = (3s \cos t, 3s \sin t, 4t). \]
Find a formula for the area of \(\mathbf{r}(U)\) in terms of integrals. This is called a helicoid. Here is a picture of it.
25.2. THE TWO DIMENSIONAL AREA IN $\mathbb{R}^3$

The area element is

$$| (3 \cos(t), 3 \sin(t), 0) \times (-3 \sin(t), 3 \cos(t), 4)| \, dsdt = \sqrt{144 + (9 (\cos^2 t) s + 9 \sin^2 t)^2} \, dsdt$$

Therefore, the area is given by the double integral,

$$\int_{-5}^{5} \int_{0}^{\pi} \sqrt{144 + (9 (\cos^2 t) s + 9 \sin^2 t)^2} \, dsdt$$

You can define the center of mass and density of a surface in exactly the same way as was done before.

**Definition 25.2.10** Let $S$ be a surface with area (volume) element $dS$. The **density with respect to area** is a function which integrated gives the mass. Thus if $\delta(x)$ is the density, the mass of $S$ is

$$\int_S \delta(x,y,z) \, dS.$$  

The **center of mass** is defined exactly as before.

$$x_c = \frac{\int_S \delta(x,y,z) x \, dS}{\int_S \delta(x,y,z) \, dS}, \quad y_c = \frac{\int_S \delta(x,y,z) y \, dS}{\int_S \delta(x,y,z) \, dS}, \quad z_c = \frac{\int_S \delta(x,y,z) z \, dS}{\int_S \delta(x,y,z) \, dS}.$$  

There is no new thing here. You simply are integrating over a surface rather than a volume. Of course you must put the variables, $x,y,z$ as well as $dS$ in terms of the parameters used to compute the integrals.

**Example 25.2.11** The surface is given by $(x,y,z) = (\sin \phi \cos \theta, \sin \phi \sin \theta, \cos \phi)$ where $(\phi, \theta) \in (0, \pi) \times (0,2\pi)$. Thus the surface is the surface of a sphere of radius 1. Review spherical coordinates at this time if this is not obvious to you. Suppose the density of a point on this surface corresponding to $(\phi, \theta)$ is $\sin^2 \phi$. That is, the density is equal to the square of the distance to the $z$ axis. Find the total mass and the center of mass of this surface.

First find the area element. This equals

$$dS = \sqrt{(\cos^2 \phi \cos \theta)^2 + (\sin^2 \phi \sin \theta)^2 + (\cos \phi \cos^2 \theta \sin \phi + \cos \phi \sin^2 \theta \sin \phi)^2} \, d\theta d\phi$$

$$= \sin(\phi) \, d\theta d\phi$$
To find the total mass you must integrate this area element times the density. Thus the total mass is
\[
\int_0^\pi \int_0^{2\pi} (\sin^2(\phi)) \sin(\phi) \, d\theta d\phi = \frac{8}{3} \pi.
\]

Next you want to find the center of mass. By symmetry, it should be at the origin. As before, the center of mass does not need to be in the surface just as it did not need to be in the three dimensional shape. Now on the surface, \(x = \sin \phi \cos \theta, y = \sin \phi \sin \theta,\) and \(z = \cos \phi.\) Consider the formulas for this.

\[
x_c = \frac{\int_0^\pi \int_0^{2\pi} (\sin(\phi) \cos(\theta)) (\sin^2(\phi)) \sin(\phi) \, d\theta d\phi}{\left(\frac{8}{3} \pi\right)} = 0,
\]

\[
y_c = \frac{\int_0^\pi \int_0^{2\pi} (\sin(\phi) \sin(\theta)) (\sin^2(\phi)) \sin(\phi) \, d\theta d\phi}{\left(\frac{8}{3} \pi\right)} = 0,
\]

\[
z_c = \frac{\int_0^\pi \int_0^{2\pi} (\cos(\phi)) (\sin^2(\phi)) \sin(\phi) \, d\theta d\phi}{\left(\frac{8}{3} \pi\right)} = 0.
\]

**Example 25.2.12** In the above example suppose \(\delta(x, y, z) = z + 1.\) What is the mass and center of mass?

The total mass is
\[
\int_0^\pi \int_0^{2\pi} (1 + \cos(\phi)) \sin(\phi) \, d\theta d\phi = 4\pi.
\]

Next, the center of mass is given by

\[
x_c = \frac{\int_0^\pi \int_0^{2\pi} (\sin(\phi) \cos(\theta)) (1 + \cos(\phi)) \sin(\phi) \, d\theta d\phi}{4\pi} = 0,
\]

\[
y_c = \frac{\int_0^\pi \int_0^{2\pi} (\sin(\phi) \sin(\theta)) (1 + \cos(\phi)) \sin(\phi) \, d\theta d\phi}{4\pi} = 0,
\]

\[
z_c = \frac{\int_0^\pi \int_0^{2\pi} (\cos(\phi)) (1 + \cos(\phi)) \sin(\phi) \, d\theta d\phi}{4\pi} = \frac{1}{3}.
\]

### 25.2.1 Surfaces Of The Form \(z = f(x, y)\)

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 \(r(x, y) = (x, y, f(x, y))^T\) such that \((x, y) \in U.\) Then

\[
r_x = \begin{pmatrix} 1 \\ 0 \\ f_x \end{pmatrix}, \quad r_y = \begin{pmatrix} 0 \\ 1 \\ f_y \end{pmatrix}
\]

and

\[|r_x \times r_y| = \sqrt{1 + f_y^2 + f_x^2} \]

so the area element is

\[\sqrt{1 + f_y^2 + f_x^2} \, dx \, dy.\]

When the surface of interest comes in this simple form, people generally use this area element directly rather than worrying about a parametrization and taking cross products.
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 \(\frac{1}{\sqrt{1 + f_x^2 + f_y^2}} (-f_x, -f_y, 1)\).

Now consider the following picture.

\[
\begin{array}{c}
n \\
\theta \\
\hline
k
\end{array}
\]

In this picture, you are looking at a chunk of area \( dS \) on the surface seen on edge and so it seems reasonable to expect to have \( dx \, dy = dS \cos \theta \). 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, \( dS = \sqrt{1 + f_x^2 + f_y^2} \, dx \, dy \) as claimed.

**Example 25.2.13** Let \( z = \sqrt{x^2 + y^2} \) where \((x, y) \in U \) for \(U = \{(x, y) : x^2 + y^2 \leq 4\} \) Find
\[
\int S h \, dS
\]
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 \( dS = \sqrt{2} \, dx \, dy \). 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,
\[
\int_S h \, dS = \int_U \left( x + \sqrt{x^2 + y^2} \right) \sqrt{2} \, dA
\]
\[
= \sqrt{2} \int_0^{2\pi} \int_0^2 (r \cos \theta + r) \, r 
\]
\[
\int_0^2 (r \cos \theta + r) \, r 
\]
\[
= \frac{16}{3} \sqrt{2} \pi.
\]

One other issue is worth mentioning. Suppose \( \mathbf{r}_1 : U_1 \to \mathbb{R}^3 \) where \( U_1 \) are sets in \( \mathbb{R}^2 \) and suppose \( \mathbf{r}_1(U_1) \) intersects \( \mathbf{r}_2(U_2) \) along \( C \) where \( C = \mathbf{h}(V) \) for \( V \subseteq \mathbb{R}^1 \). Then define integrals and areas over \( \mathbf{r}_1(U_1) \cup \mathbf{r}_2(U_2) \) as follows.
\[
\int_{\mathbf{r}_1(U_1) \cup \mathbf{r}_2(U_2)} g \, dS = \int_{\mathbf{r}_1(U_1)} g \, dS + \int_{\mathbf{r}_2(U_2)} g \, dS.
\]
Admittedly, the set $C$ gets added in twice but this doesn’t matter because its 2 dimensional volume equals zero and therefore, the integrals over this set will also be zero.

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 $|r(x) - r(y)| \leq K|x - y|$. 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. There are many very interesting issues here which can keep you fascinated for years.

### 25.3 Flux

Imagine a surface, $S$ which is fixed in space and let $\mathbf{v}$ be a vector field representing the velocity of a fluid flowing through this surface. It is reasonable to ask how fast the fluid crosses the surface in terms of units of mass per units of time. This is expressed in terms of the surface integral,

$$
\int_S \rho \mathbf{v} \cdot \mathbf{n} \, dA
$$

where $\rho$ is the density and $\mathbf{n}$ is the normal vector to the surface in the direction in which the crossing is taking place. The vector field, $\rho \mathbf{v}$ is called the flux. To get the rate of transfer of mass across the surface, you take the dot product of the flux with the appropriate unit normal vector and integrate this over the surface. People also speak of heat flux. In general, when they speak of flux, they mean the thing you dot with a unit normal vector and integrate to find the rate at which something crosses a surface. A little later, this idea will be explored much more when the divergence theorem is established. It is a very important idea. You should think about the physical reasons the flux of such a fluid is given as above. Why do you use the unit normal for example? Why not some normal which has different length? Why do you need to take the dot product with the normal? In general situations, people assume formulas about the flux in terms of other quantities such as temperature or concentration. I will mention some later at a convenient place.

Here is an example of a flux integral.

**Example 25.3.1** Let $\mathbf{F} = (x, y, z)$ and let $S$ be the curved surface which comes from the intersection of the plane $z = x$ with the paraboloid $z = x^2 + y^2$. Find an iterated integral for the flux integral

$$
\int_S \mathbf{F} \cdot \mathbf{n} \, dS
$$

where $\mathbf{n}$ is the field of unit normals which has negative $z$ component.
25.3. FLUX

First find $ds$. This equals $\sqrt{1+4x^2+4y^2}dx\,dy$. Now consider the circle in the plane which is below the surface. The boundary of this circle is $(x, y)$ such that the two surfaces intersect. Thus

$$x^2 - x + y^2 = 0$$

and so

$$\left(x - \frac{1}{2}\right)^2 + y^2 = \frac{1}{4}$$

which happens to be a circle of radius $\frac{1}{2}$ having center at $\left(\frac{1}{2}, 0\right)$. In polar coordinates this is

$$r = \cos \theta, \quad \theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$$

Now what is $\mathbf{F} \cdot \mathbf{n}$? \( \mathbf{n} = (2x, 2y, -1) \frac{1}{\sqrt{1+4(x^2+y^2)}} \) and so

$$\mathbf{F} \cdot \mathbf{n} = (2x, 2y, -1) \frac{1}{\sqrt{1+4(x^2+y^2)}} \cdot (x, y, z)$$

$$= \frac{2x^2 + 2y^2 - z}{\sqrt{(1+4x^2+y^2)}} = \frac{2x^2 + 2y^2 - (x^2 + y^2)}{\sqrt{(1+4x^2+y^2)}}$$

$$= \frac{x^2 + y^2}{\sqrt{(1+4x^2+y^2)}}$$

Then you need to integrate

$$\frac{x^2 + y^2}{\sqrt{(1+4x^2+y^2)}} \sqrt{1+4x^2+4y^2} dx\,dy = (x^2 + y^2) \, dx\,dy$$

Denoting this circle by $D$, You can now finish this problem. Use polar coordinates.

$$\int_{-\pi/2}^{\pi/2} \int_{0}^{\cos \theta} r^2 r \, dr \, d\theta = \frac{3}{32} \pi$$

There is a “shortcut” when the surface comes to you in parametric form. Suppose $S$ is a smooth surface and it has parametrization $\mathbf{r}(s, t)$ where the parameter domain is $D \subseteq \mathbb{R}^2$, $\mathbf{r}_s \times \mathbf{r}_t \neq 0$. Also suppose $\mathbf{F}$ is a vector field defined on $S$ and the unit normal to $S$ has the same direction as $\mathbf{r}_s \times \mathbf{r}_t$. Recall that $\mathbf{r}_s \times \mathbf{r}_t$ is a normal to $S$. Then from the above, the flux integral $\int_S \mathbf{F} \cdot \mathbf{n} dS$ is of the form

$$\int_S \mathbf{F} \cdot \mathbf{n} dS = \int_D \mathbf{F}(\mathbf{r}(s, t)) \cdot \frac{\mathbf{r}_s \times \mathbf{r}_t}{|\mathbf{r}_s \times \mathbf{r}_t|} |\mathbf{r}_s \times \mathbf{r}_t| \, ds \, dt$$

Notice how $|\mathbf{r}_s \times \mathbf{r}_t|$ cancels. Thus the flux integral reduces to

$$\int_D \mathbf{F}(\mathbf{r}(s, t)) \cdot \mathbf{r}_s \times \mathbf{r}_t ds \, dt$$

**Example 25.3.2** Do the above example this way. Recall $\mathbf{F} = (x, y, z)$.

You have $z = x^2 + y^2$ and $(x, y)$ is in $D : (x - \frac{1}{2})^2 + y^2 \leq \frac{1}{4}$. Then a parametrization is

$$\begin{pmatrix} x \\ y \\ x^2 + y^2 \end{pmatrix}$$
CHAPTER 25. THE INTEGRAL ON TWO DIMENSIONAL SURFACES IN $\mathbb{R}^3$

Then you need $\mathbf{r}_x(x,y) \times \mathbf{r}_y(x,y)$ is

$$\begin{pmatrix} 1 \\ 0 \\ 2x \end{pmatrix} \times \begin{pmatrix} 0 \\ 1 \\ 2y \end{pmatrix} = \begin{pmatrix} -2x \\ -2y \\ 1 \end{pmatrix}$$

Now this is not the one we want! This is because it points up rather than down. The one we want is

$$\mathbf{r}_y(x,y) \times \mathbf{r}_x(x,y) = \begin{pmatrix} 2x \\ 2y \\ -1 \end{pmatrix}$$

Then the integral is

$$\int_D \begin{pmatrix} x \\ y \\ x^2 + y^2 \end{pmatrix} \cdot \begin{pmatrix} 2x \\ 2y \\ -1 \end{pmatrix} dxdy = \int_D x^2 + y^2 dxdy$$

This is the same as earlier. You change to polar coordinates and work the integral. Thus

$$\int_D x^2 + y^2 dxdy = \int_{\pi/2}^{\pi/2} \int_0^{\cos \theta} r^2 r dr d\theta = \frac{3}{32} \pi.$$

Example 25.3.3 Let $\mathbf{F} = (x, -x, z)$ and find $\int_S \mathbf{F} \cdot \mathbf{n} dS$ where $S$ is the top half of the unit sphere centered at $(0,0,0)$ and $\mathbf{n}$ points upward.

Lets parametrize this.

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \sin(\phi) \cos(\theta) \\ \sin(\phi) \sin(\theta) \\ \cos(\phi) \end{pmatrix}, \phi \in [0, \pi/2], \theta \in [0, 2\pi]$$

consider $\mathbf{r}_\phi \times \mathbf{r}_\theta$

$$\begin{pmatrix} \cos \theta \cos \phi \\ \cos \phi \sin \theta \\ -\sin \phi \end{pmatrix} \times \begin{pmatrix} -\sin \theta \sin \phi \\ \cos \theta \sin \phi \\ 0 \end{pmatrix} = \begin{pmatrix} \cos \theta \sin^2 \phi \\ \sin \theta \sin^2 \phi \\ \cos \phi \sin \phi \end{pmatrix}$$

Is this the one we want? Yes. It points up. If it had not done so, we would have multiplied by $-1$.

$$\int_{\pi/2}^{\pi/2} \int_0^{2\pi} \begin{pmatrix} \sin(\phi) \cos(\theta) \\ -\sin(\phi) \cos(\theta) \\ \cos(\phi) \end{pmatrix} \cdot \begin{pmatrix} \cos \theta \sin^2 \phi \\ \sin \theta \sin^2 \phi \\ \cos \phi \sin \phi \end{pmatrix} d\theta d\phi =$$

$$\int_0^{\pi/2} \int_0^{2\pi} ( \cos^2 \theta \sin^3 \phi - \sin \theta \cos \theta \sin^3 \phi + \cos^2 \phi \sin \phi ) d\theta d\phi = \frac{4}{3} \pi$$

Procedure for finding the flux integral
Procedure 25.3.4 Finding the flux of a vector field $\mathbf{F}$ across a surface $S$.

$$\int_S \mathbf{F} \cdot \mathbf{n} \, dS$$

1. Parametrize the surface. That is, find a parameter domain $D \subseteq \mathbb{R}^2$ and a function $\mathbf{r} : D \rightarrow S = \mathbf{r}(D)$.

2. Pick either $\mathbf{r}_t \times \mathbf{r}_s$ or $\mathbf{r}_s \times \mathbf{r}_t$ whichever one has the same direction as $\mathbf{n}$.

3. If the choice in 2. is $\mathbf{r}_t \times \mathbf{r}_s$, the flux integral is $\int_D \mathbf{F}(\mathbf{r}(t,s)) \cdot (\mathbf{r}_t \times \mathbf{r}_s) \, dsdt$ and if the choice in 2. is $\mathbf{r}_s \times \mathbf{r}_t$, then the flux integral is $\int_D \mathbf{F}(\mathbf{r}(t,s)) \cdot (\mathbf{r}_s \times \mathbf{r}_t) \, dsdt$.

You can also just follow what the symbols mean. That is, find $\mathbf{F} \cdot \mathbf{n}$ on the surface in terms of two variables, find $dS$ in terms of these variables and reduce directly to an integral over a region in the plane.

To emphasize the above procedure again, here is another example.

Example 25.3.5 Find $\int_S \mathbf{F} \cdot \mathbf{n} \, dS$ where $\mathbf{F}(x,y,z) = (x,z,y)$ and $\mathbf{n}$ is the unit normal to $S$ having negative $z$ coordinate and $S$ is the part of the plane $2z + 2y + x = 3$ which lies in the cylinder $x^2 + y^2 = 1$.

Let's find $\mathbf{n}$. It equals $\frac{(1,2,2)}{\sqrt{1+4+4}} = \frac{(1,2,2)}{3}$. The minus sign is there because the normal is given to point down. Next find $dS$. It will not involve something as simple as the norm of $\mathbf{n}$ as it does when $\mathbf{n}$ is given by an expression like $\mathbf{r}_s \times \mathbf{r}_t$. Solve for $z$.

$$z = \frac{3}{2} - y - \frac{1}{2}x$$

Then

$$dS = \sqrt{1 + 1 + \frac{1}{4}} \, dx \, dy = \frac{3}{2} \, dx \, dy$$

It follows that $\mathbf{F} \cdot \mathbf{n} = (x,z,y) \cdot \left( \frac{-1,-2,-2}{3} \right) = \frac{1}{3}x - \frac{2}{3}y - \frac{2}{3}z$. Now we must replace $z$ with what it equals on this surface. Thus

$$\mathbf{F} \cdot \mathbf{n} = - \left( \frac{1}{3}x + \frac{2}{3}y + \frac{2}{3}z \right) = - \left( \frac{1}{3}x + \frac{2}{3}y + \frac{2}{3} \left( \frac{3}{2} - y - \frac{1}{2}x \right) \right) = -1$$

where $(x,y)$ is in $D$ the unit disk centered at $(0,0)$. It follows that

$$\int_S \mathbf{F} \cdot \mathbf{n} \, dS = \int_D (-1) \frac{3}{2} \, dx \, dy = -\frac{3}{2} \pi$$

because the area of $D$ is $\pi$.

Let's do this the other way by writing a parametrization. Such a parametrization is

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} s \\ t \\ \frac{3}{2} - t - \frac{1}{2}s \end{pmatrix}, (s,t) \text{ in the unit disk.}$$

Then

$$\frac{\partial}{\partial s} \begin{pmatrix} s \\ t \\ \frac{3}{2} - t - \frac{1}{2}s \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -\frac{1}{2} \end{pmatrix}, \frac{\partial}{\partial t} \begin{pmatrix} s \\ t \\ \frac{3}{2} - s - \frac{1}{2}s \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$
Then we must take the cross product of these.

\[
\begin{pmatrix} 1 \\ 0 \\ -1/2 \end{pmatrix} \times \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1 \\ 1 \end{pmatrix}
\]

This is not the one we want because it is given that the component of \( z \) is negative. Thus the one we want is

\[
\begin{pmatrix} -1/2 \\ -1 \\ -1 \end{pmatrix}
\]

Now the flux integral is just

\[
\int_D \left( \begin{pmatrix} \frac{s}{2} - \frac{1}{2}s \\ s - \frac{1}{2}s \\ t \end{pmatrix} \cdot \begin{pmatrix} -1/2 \\ -1 \\ -1 \end{pmatrix} \right) dsdt
\]

\[
= \int_D s - t - \frac{3}{2} dt ds = -\frac{3}{2} \pi
\]

because the \( s, t \) integrate to 0 on \( D \).

You can do the flux integral either way, just substituting in for \( \mathbf{F} \cdot \mathbf{n} \) and \( dS \) or following a ritual of finding a parametrization dotting with the cross product of two vector valued functions and integrating over the parameter domain.

### 25.4 Exercises With Answers

1. Find a parametrization for the intersection of the planes \( x+y+2z = -3 \) and \( 2x-y+z = -4 \).
   
   Answer:
   \((x, y, z) = (-t - 7/3, -t - 2/3, t)\)

2. Find a parametrization 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 parametrization 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 parametrization 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 parametrization 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 parametrization 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 parametrization 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. \(y = -2x^2 - \frac{1}{2} + \frac{3}{4}x, z = -\frac{1}{4}x + \frac{3}{2} + 2x^2\). Therefore, the parametrization is \((x, y, z) = (t, -2t^2 - \frac{1}{2} + \frac{3}{2}t, -\frac{1}{4}t + \frac{3}{2} + 2t^2)\).

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 parametrization, \(x = 4\cos v, y = 4\sin v\) 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:

   ![Diagram of a side view of a cylinder](image)

   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 \(A = \int_{-\pi/2}^{3\pi/2} \int_0^{4+4\sin v} 4\,du\,dv = 32\pi\).

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 parametrization of this surface is \(x = u, y = v, z = \frac{1}{3}\sqrt{(u^2 + v^2)}\) where \((u, v) \in D\), a disk centered at the origin having radius \(ha\). Therefore, the volume is just \(\int_D \sqrt{1 + z_u^2 + z_v^2} \, dA = \int_{-ha}^{ha} \int_{-\sqrt{(9h^2 - u^2)}}^{\sqrt{(9h^2 - u^2)}} \frac{1}{3}\sqrt{10} \, dv \, du = 3\pi h^2\sqrt{10}\).

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
\[ |\mathbf{f}_u \times \mathbf{f}_v| = \begin{vmatrix} 0 & -2 \sin v & 0 \\ 0 & 2 \cos v & 0 \\ 1 & 0 & 0 \end{vmatrix} = 2. \]
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
\[
\int_{\theta=0}^{\theta=2\pi} \int_{r=0}^{r=2+\cos \theta} r \, dr \, d\theta = \frac{9}{2} \pi.
\]

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
\[
\int_R \sqrt{1 + 4x^2h^2 + 4y^2h^2} \, dA.
\]
Changing to polar coordinates, this becomes
\[
\int_0^{2\pi} \int_0^h (\sqrt{1 + 4h^2r^2}) \, r \, dr \, d\theta = \frac{\pi}{6h^2} \left( (1 + 4h^2)^{3/2} - 1 \right).
\]

11. Evaluate \( \int_S (1 + x) \, dA \) where \( S \) is the part of the plane \( 2x + 3y + 3z = 18 \) which is in the first octant.

Answer:
\[
\int_0^6 \int_{-6-x}^{6-x} (1 + x) \frac{1}{2} \sqrt{22} \, dy \, dx = 28 \sqrt{22}
\]

12. Evaluate \( \int_S (1 + x) \, dA \) 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 \( \int_S (1 + x) \, dA \) 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 \( x > 0 \), \( \theta \in [-\frac{\pi}{2}, \frac{\pi}{2}] \) 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
\[
\int_0^\pi \int_{-\pi/2}^{\pi/2} (1 + 4 \sin \phi \cos \theta) 16 \sin \phi \, d\theta \, d\phi = 96\pi
\]
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:
\[
|f_\theta \times f_\alpha| = \begin{vmatrix}
-\sin (\theta) (2 + \cos \alpha) & -\cos \theta \sin \alpha \\
-\cos (\theta) (2 + \cos \alpha) & \sin \theta \sin \alpha \\
0 & \cos \alpha
\end{vmatrix} = (4 + 4 \cos \alpha + \cos^2 \alpha)^{1/2}
\]

and so the area element is
\[(4 + 4 \cos \alpha + \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha.
\]

Therefore, the area is
\[
\int_0^{2\pi} \int_0^{2\pi} (4 + 4 \cos \alpha + \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha = \int_0^{2\pi} \int_0^{2\pi} (2 + \cos \alpha) \, d\theta \, d\alpha = 8\pi^2.
\]

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 \(\int_{f([0,2\pi]\times[0,2\pi])} h \, dA\).

Answer:
\[
|f_\theta \times f_\alpha| = \begin{vmatrix}
-\sin (\theta) (4 + 2 \cos \alpha) & -2 \cos \theta \sin \alpha \\
-\cos (\theta) (4 + 2 \cos \alpha) & 2 \sin \theta \sin \alpha \\
0 & 2 \cos \alpha
\end{vmatrix} = (64 + 64 \cos \alpha + 16 \cos^2 \alpha)^{1/2}
\]

and so the area element is
\[(64 + 64 \cos \alpha + 16 \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha.
\]

Therefore, the desired integral is
\[
\int_0^{2\pi} \int_0^{2\pi} (\cos \alpha) (64 + 64 \cos \alpha + 16 \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha
\]
\[
= \int_0^{2\pi} \int_0^{2\pi} (\cos \alpha) (8 + 4 \cos \alpha) \, d\theta \, d\alpha = 8\pi^2
\]
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:
The area element is \((9 + 6 \cos \alpha + \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha\).
Therefore, the desired integral is
\[
\int_0^{2\pi} \int_0^{2\pi} \left(\cos^2 \theta\right) (9 + 6 \cos \alpha + \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha
= \int_0^{2\pi} \int_0^{2\pi} \left(\cos^2 \theta\right) (3 + \cos \alpha) \, d\theta \, d\alpha = 6\pi^2
\]
17. For \((\theta, \alpha) \in [0, 25] \times [0, 2\pi]\), let
\[
f(\theta, \alpha) \equiv (\cos \theta (4 + 2 \cos \alpha), -\cos \beta \sin \theta (4 + 2 \cos \alpha), 2 \sin \beta \sin \theta (2 + \cos \alpha) + \cos \beta \sin \alpha)^T.
\]
Find a double integral which gives the area of \(f([0, 25] \times [0, 2\pi])\).
Answer:
In this case, the area element is
\[(68 + 64 \cos \alpha + 12 \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha\]
and so the surface 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 \(f(\theta, \alpha) \equiv (\cos \theta (2 + \cos \alpha), -\cos \beta \sin \theta (2 + \cos \alpha) + \sin \beta \sin \alpha, 
\sin \beta \sin \theta (2 + \cos \alpha) + \cos \beta \sin \alpha)^T\).
Find a double integral which gives the area of \(f([0, 2\pi] \times [0, 2\pi])\).
Answer:
After many computations, the area element is \((4 + 4 \cos \alpha + \cos^2 \alpha)^{1/2} \, d\theta \, d\alpha\). Therefore, the area is \(\int_0^{2\pi} \int_0^{2\pi} (2 + \cos \alpha) \, d\theta \, d\alpha = 8\pi^2\).
Part IX

The Divergence Theorem And Stoke’s Theorem
Chapter 26

Divergence And Curl

26.1 Divergence Of A Vector Field

Here the important concepts of divergence is defined.

**Definition 26.1.1** Let \( f : U \to \mathbb{R}^p \) 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 \( i \). 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.

\[
f \cdot \nabla (g) \equiv f_1(x) \frac{\partial g(x)}{\partial x_1} + f_2(x) \frac{\partial g(x)}{\partial x_2} + \cdots + f_p(x) \frac{\partial g(x)}{\partial x_p}.
\]

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 a coordinate free definition of div is presented. For now, everything is defined in terms of a given Cartesian coordinate system. The divergence has profound physical significance and this will be discussed later. For now it is important to understand how to find it. Be sure you understand that for \( f \) a vector field, \( \text{div} f \) is a vector field described earlier.

**Example 26.1.2** Let \( f(x) = xyi + (z - y)j + (\sin(x) + z)k \). Find \( \text{div} f \)
First the divergence of \( f \) is
\[
\frac{\partial (xy)}{\partial x} + \frac{\partial (z - y)}{\partial y} + \frac{\partial (\sin(x) + z)}{\partial z} = y + (-1) + 1 = y.
\]

### 26.2 Curl Of A Vector Field

Here the important concepts of curl is defined.

**Definition 26.2.1** Let \( f : U \to \mathbb{R}^3 \) for \( U \subseteq \mathbb{R}^3 \) denote a vector field. The **curl** of the vector field yields another vector field and it is defined as follows.

\[
(\text{curl} (f)(x))_i \equiv (\nabla \times f)(x)_i \equiv \varepsilon_{ijk} \frac{\partial}{\partial x_j} f_k(x)
\]

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.

\[
\begin{vmatrix}
  i & j & k \\
  \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\
  f_1(x, y, z) & f_2(x, y, z) & f_3(x, y, z)
\end{vmatrix}.
\]

Note the similarity with the cross product. Sometimes the curl is called rot. (Short for rotation not decay.)

This definition is in terms of a given coordinate system but later coordinate free definitions of the curl is presented. For now, everything is defined in terms of a given Cartesian coordinate system. The curl has profound physical significance and this will be discussed later. For now it is important to understand how to find it. Be sure you understand that for \( f \) a vector field, \( \text{curl} f \) is another vector field.

**Example 26.2.2** Let \( f(x) = xyi + (z - y)j + (\sin(x) + z)k \). Find \( \text{curl} f \).

\( \text{curl} f \) is obtained by evaluating

\[
\begin{vmatrix}
  i & j & k \\
  \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\
  xy & z - y & \sin(x) + z
\end{vmatrix} = 
\]

\[
i \left( \frac{\partial}{\partial y} (\sin(x) + z) - \frac{\partial}{\partial z} (z - y) \right) - j \left( \frac{\partial}{\partial x} (\sin(x) + z) - \frac{\partial}{\partial z} (xy) \right) + 
\]

\[
k \left( \frac{\partial}{\partial x} (z - y) - \frac{\partial}{\partial y} (xy) \right) = -i - \cos(x) j - xk.
\]
Chapter 27

The Divergence Theorem

Why does anyone care about the divergence of a vector field? The answer is contained in this section. In short, it is because of the divergence theorem which relates the flux over the boundary to a volume integral of the divergence. It is also called Gauss’s theorem.

Definition 27.0.3 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 27.0.4** 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 $\mathbf{n} = (n_x, n_y, n_z)$ be the unit exterior normal to $\partial V$. Then

$$\int_V \frac{\partial F}{\partial z} (x, y, z) \, dV = \int_{\partial V} Fn_z \, dA.$$  

**Proof:** From the fundamental theorem of calculus,

$$\int_V \frac{\partial F}{\partial z} (x, y, z) \, dV = \int_D \int_{\phi(x,y)}^{\psi(x,y)} \frac{\partial F}{\partial z} (x, y, z) \, dz \, dx \, dy$$  

$$= \int_D [F (x, y, \psi (x,y)) - F (x, y, \phi (x,y))] \, dx \, dy$$  

Now the unit exterior normal on the top of $V$, the surface $(x, y, \psi (x,y))$ is

$$\frac{1}{\sqrt{\psi_x^2 + \psi_y^2 + 1}} (-\psi_x, -\psi_y, 1).$$

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
\[
\frac{1}{\sqrt{\phi_x^2 + \phi_y^2 + 1}} (\phi_x, \phi_y, -1)
\]
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 (27.1) is of the form,
\[
\int_D F(x, y, \psi(x, y)) \left( \frac{1}{\sqrt{\psi_x^2 + \psi_y^2 + 1}} \right) \sqrt{\psi_x^2 + \psi_y^2 + 1} \, dx \, dy + \int_{\text{Lateral surface}} F n_z \, dA,
\]
the last term equaling zero because on the lateral surface, \( n_z = 0 \). Therefore, this reduces to \( \int_{\partial V} F n_z \, dA \) as claimed.

The following corollary is entirely similar to the above.

Corollary 27.0.5 If \( V \) is cylindrical in the \( y \) direction, then
\[
\int_V \frac{\partial F}{\partial y} \, dV = \int_{\partial V} F n_y \, dA
\]
and if \( V \) is cylindrical in the \( x \) direction, then
\[
\int_V \frac{\partial F}{\partial x} \, dV = \int_{\partial V} F n_x \, dA
\]

With this corollary, here is a proof of the divergence theorem.

Theorem 27.0.6 Let \( V \) be cylindrical in each of the coordinate directions and let \( F \) be a \( C^1 \) vector field defined on \( V \). Then
\[
\int_V \nabla \cdot F \, dV = \int_{\partial V} F \cdot n \, dA.
\]
CHAPTER 27. THE DIVERGENCE THEOREM

**Proof:** From the above lemma and corollary,

\[
\int_V \nabla \cdot \mathbf{F} \, dV = \int_V \left( \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z} \right) \, dV
\]

\[
= \int_{\partial V} (F_1 n_x + F_2 n_y + F_3 n_z) \, dA
\]

\[
= \int_{\partial V} \mathbf{F} \cdot \mathbf{n} \, dA.
\]

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.

![Picture of two overlapping regions](image)

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 it can be used with confidence.

**Example 27.0.7** Let \( V = [0,1] \times [0,1] \times [0,1] \). That is, \( V \) is the cube in the first octant having the lower left corner at \((0,0,0)\) and the sides of length 1. Let \( \mathbf{F}(x,y,z) = xi + yj + zk \). Find the flux integral in which \( \mathbf{n} \) is the unit exterior normal.

\[
\int_{\partial V} \mathbf{F} \cdot \mathbf{n} \, dS
\]

You can certainly inflict much suffering on yourself by breaking the surface up into 6 pieces corresponding to the 6 sides of the cube, finding a parametrization for each face and adding up the appropriate flux integrals. For example, \( \mathbf{n} = \mathbf{k} \) on the top face and \( \mathbf{n} = -\mathbf{k} \) on the bottom face. On the top face, a parametrization is \((x,y,1) : (x,y) \in [0,1] \times [0,1] \). The area element is just \( dx \, dy \). It isn’t really all that hard to do it this way but it is much easier to use the divergence theorem. The above integral equals

\[
\int_V \text{div} (\mathbf{F}) \, dV = \int_V 3 \, dV = 3.
\]

**Example 27.0.8** This time, let \( V \) be the unit ball, \( \{(x,y,z) : x^2 + y^2 + z^2 \leq 1\} \) and let \( \mathbf{F}(x,y,z) = x^2 i + yj + (z - 1)k \). Find

\[
\int_{\partial V} \mathbf{F} \cdot \mathbf{n} \, dS.
\]
As in the above you could do this by brute force. A parametrization of the $\partial V$ is obtained as
\[ x = \sin \phi \cos \theta, \; y = \sin \phi \sin \theta, \; z = \cos \phi \]
where $(\phi, \theta) \in (0, \pi) \times (0, 2\pi]$. Now this does not include all the ball but it includes all but the point at the top and at the bottom. As far as the flux integral is concerned these points contribute nothing to the integral so you can neglect them. Then you can grind away and get the flux integral which is desired. However, it is so much easier to use the divergence theorem! Using spherical coordinates,
\[
\int_{\partial V} \mathbf{F} \cdot \mathbf{n} dS = \int_V \text{div} (\mathbf{F}) dV = \int_V (2x + 1 + 1) dV
\]
\[
= \int_0^\pi \int_0^{2\pi} \int_0^1 (2 + 2\rho \sin (\phi) \cos \theta) \rho^2 \sin (\phi) d\rho d\theta d\phi = \frac{8}{3} \pi
\]

**Example 27.0.9** Suppose $V$ is an open set in $\mathbb{R}^3$ for which the divergence theorem holds. Let $\mathbf{F}(x, y, z) = xi + yj + zk$. Then show
\[
\int_{\partial V} \mathbf{F} \cdot \mathbf{n} dS = 3 \times \text{volume}(V).
\]

This follows from the divergence theorem.
\[
\int_{\partial V} \mathbf{F} \cdot \mathbf{n} dS = \int_V \text{div} (\mathbf{F}) dV = 3 \int_V dV = 3 \times \text{volume}(V).
\]

The message of the divergence theorem is the relation between the volume integral and an area integral. This is the exciting thing about this marvelous theorem. It is not its utility as a method for evaluations of boring problems. This will be shown in the examples of its use which follow.

### 27.0.1 Green’s Theorem, A Review

You can prove Green’s theorem from the divergence theorem.

**Theorem 27.0.10** (Green’s Theorem) Let $U$ be an open set in the plane and let $\partial U$ be piecewise smooth and let $\mathbf{F}(x, y) = (P(x, y), Q(x, y))$ be a $C^1$ vector field defined near $U$. Then it is often the case that
\[
\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} = \int_U \left( \frac{\partial Q}{\partial x} (x, y) - \frac{\partial P}{\partial y} (x, y) \right) dA.
\]

**Proof:** Suppose the divergence theorem holds for $U$. Consider the following picture.

---

For a general version see the advanced calculus book by Apostol. The general versions involve the concept of a rectifiable Jordan curve.
Since it is assumed that motion around \( U \) is counter clockwise, the tangent vector, \((x',y')\) is as shown. Now the unit exterior normal is either
\[
\frac{1}{\sqrt{(x')^2 + (y')^2}} (-y', x')
\]
or
\[
\frac{1}{\sqrt{(x')^2 + (y')^2}} (y', -x')
\]
Again, the counter clockwise motion shows the correct unit exterior normal is the second of the above. To see this note that since the area should be on the left as you walk around the edge, you need to have the unit normal point in the direction of \((x', y', 0) \times k\) which equals \((y', -x', 0)\). Now let \( F \ (x, y) = (Q \ (x, y), -P \ (x, y)) \). Also note the area element on \( \partial U \) is \( \sqrt{(x')^2 + (y')^2} \ dt \). Suppose the boundary of \( U \) consists of \( m \) smooth curves, the \( i^{th} \) of which is parametrized by \((x_i, y_i)\) with the parameter, \( t \in [a_i, b_i] \). Then by the divergence theorem,
\[
\int_U (Q_x - P_y) \ dA = \int_{\partial U} \text{div} (F) \ dA = \int_{\partial U} F \cdot n \ dS
\]

\[
= \sum_{i=1}^{m} \int_{a_i}^{b_i} \left( Q \ (x_i (t), y_i (t)), -P \ (x_i (t), y_i (t)) \right) \cdot \frac{1}{\sqrt{(x'_i)^2 + (y'_i)^2}} \ \frac{dS}{\sqrt{(x'_i)^2 + (y'_i)^2}} \ dt
\]

\[
= \sum_{i=1}^{m} \int_{a_i}^{b_i} \left( Q \ (x_i (t), y_i (t)), -P \ (x_i (t), y_i (t)) \right) \cdot \left( y'_i, -x'_i \right) \ dt
\]

\[
= \sum_{i=1}^{m} \int_{a_i}^{b_i} Q \ (x_i (t), y_i (t)) y'_i (t) + P \ (x_i (t), y_i (t)) x'_i (t) \ dt \equiv \int_{\partial U} P \ dx + Q \ dy
\]

This proves Green’s theorem from the divergence theorem.

### 27.0.2 Coordinate Free Concept Of Divergence, Flux Density

The divergence theorem also makes possible a coordinate free definition of the divergence.

**Theorem 27.0.11** Let \( B(x, \delta) \) be the ball centered at \( x \) having radius \( \delta \) and let \( F \) 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

\[
\text{div} \ F \ (x) = \lim_{\delta \to 0^+} \frac{1}{v(B(x, \delta))} \int_{\partial B(x, \delta)} F \cdot n \ dA. \tag{27.2}
\]

**Proof:** The divergence theorem holds for balls because they are cylindrical in every direction. Therefore,

\[
\frac{1}{v(B(x, \delta))} \int_{\partial B(x, \delta)} F \cdot n \ dA = \frac{1}{v(B(x, \delta))} \int_{B(x, \delta)} \text{div} \ F \ (y) \ dV.
\]
27.1. THE WEAK MAXIMUM PRINCIPLE

Therefore, since \( \text{div} \ F (x) \) is a constant,

\[
\left| \text{div} \ F (x) - \frac{1}{v(B(x, \delta))} \int_{B(x, \delta)} F \cdot n \, dA \right| = \left| \text{div} \ F (x) - \frac{1}{v(B(x, \delta))} \int_{B(x, \delta)} \text{div} F (y) \, dV \right|
\]

\[
\leq \frac{1}{v(B(x, \delta))} \int_{B(x, \delta)} |\text{div} F (x) - \text{div} F (y)| \, dV
\]

\[
\leq \frac{1}{v(B(x, \delta))} \int_{B(x, \delta)} \frac{\varepsilon}{2} \, dV < \varepsilon
\]

whenever \( \varepsilon \) is small enough due to the continuity of \( \text{div} F \). Since \( \varepsilon \) is arbitrary, this shows 27.2.

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.

You also see the physical significance of the divergence from this. It measures the tendency of the vector field to “diverge” from a point.

27.1 The Weak Maximum Principle

There is 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 \( U \). It is a very important result which ties in many earlier topics. Don’t read it if you are not interested.

**Theorem 27.1.1** 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 \( x_0 \in U \) such that \( u(x_0) > \max \{ u(x) : x \in \partial U \} \equiv M \). Since \( U \) is bounded, there exists \( \varepsilon > 0 \) such that

\[
u(x_0) > \max \left\{ u(x) + \varepsilon |x|^2 : x \in \partial U \right\}.
\]

Therefore, \( u(x_0) + \varepsilon |x|^2 \) also has its maximum in \( U \) because for \( \varepsilon \) small enough,

\[
u(x_0) + \varepsilon |x_0|^2 > u(x_0) > \max \left\{ u(x) + \varepsilon |x|^2 : x \in \partial U \right\}
\]

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.
27.2 Some Applications Of The Divergence Theorem

There are numerous applications of the divergence theorem. Some are listed here. You might want to read this if you are interested in applications. However, it won’t be needed for tests.

27.2.1 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

\[
- \int_{\partial B(x, \varepsilon)} p \mathbf{n} \, dA.
\]

Here \( \mathbf{n} \) is the unit exterior normal at a small piece of \( \partial B(x, \varepsilon) \) having area \( dA \). By the divergence theorem, 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

\[
\nabla p = \rho \mathbf{g}.
\]

(27.3)

It turns out that the pressure in a lake at depth \( z \) is equal to 62.5\( z \). This is easy to see from (27.3). In this case, \( \mathbf{g} = g \mathbf{k} \) 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\(^2\). Therefore, this is the mass of water at any depth. Therefore,

\[
\frac{\partial p}{\partial x} \mathbf{i} + \frac{\partial p}{\partial y} \mathbf{j} + \frac{\partial p}{\partial z} \mathbf{k} = \frac{62.5}{32} \times 32 \mathbf{k}.
\]

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 (27.3) is more interesting because it does not require \( \rho \) to be constant.

\(^2\)There is no such thing as an incompressible fluid but this doesn’t stop people from making this assumption.
27.2. SOME APPLICATIONS OF THE DIVERGENCE THEOREM

27.2.2 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,

$$\int_{\partial V} -p n \, dA = \int_V -\nabla p \, dV = \int_V \rho g \, dV$$

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 $g$ 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 Fourier's law of heat conduction because in this equation $g = -g k$.

27.2.3 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.

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,

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 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 $\int_V \rho(x, t) c(x, t) u(x, t) \, dV$. 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.

$$\frac{\partial}{\partial t} (\rho(x, t) c(x, t) u(x, t)) = \nabla \cdot (k(u, x, t) \nabla u(x, t)) + f(x, u, t). \quad (27.4)$$

Take an arbitrary $V$ for which the divergence theorem holds. Then the time rate of change of the heat in $V$ is

$$\frac{d}{dt} \int_V \rho(x, t) c(x, t) u(x, t) \, dV = \int_V \frac{\partial}{\partial t} (\rho(x, t) c(x, t) u(x, t)) \, dV$$

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,
\[
\frac{d}{dt} \int_V \rho(x,t) c(x,t) u(x,t) \, dV =
\]
\[
\int_V \frac{\partial (\rho(x,t) c(x,t) u(x,t))}{\partial t} \, dV = \int_{\partial V} -\mathbf{J} \cdot \mathbf{n} \, dA + \int_V f(x,u,t) \, dV
\]
\[
= \int_{\partial V} k \nabla (u) \cdot \mathbf{n} \, dA + \int_V f(x,u,t) \, dV = \int_V (\nabla \cdot (k \nabla (u)) + f) \, dV.
\]

Since this holds for every sample volume, \( V \) it must be the case that the above reaction diffusion equation, \((27.4)\) 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,
\[
\frac{\partial}{\partial t} u(x,t) = \frac{1}{\rho c} \nabla \cdot (k(x) \nabla u(x,t)) + f(x,t).
\]

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.

### 27.2.4 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, \( \mathbf{v}(y,t) \) is flowing through this region. Then the mass of fluid leaving \( V \) per unit time is given by the area integral, \( \int_{\partial V} \rho(y,t) \mathbf{v}(y,t) \cdot \mathbf{n} \, dA \) 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.
\[
\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = f(y,t)
\]

(27.6)

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

\[
\int_V \frac{\partial \rho(y,t)}{\partial t} \, dV = \underbrace{- \int_{\partial V} \rho(y,t) v(y,t) \cdot n \, dA}_{\text{rate at which mass enters}} + \int_V f(y,t) \, dV
\]

\[
= - \int_V (\nabla \cdot (\rho(y,t) v(y,t))) + f(y,t) \, dV.
\]

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.

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.

### 27.2.5 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 spatial coordinates and the function, \( (t,x) \to y(t,x) \) is called the motion. Thus

\[
y(0,x) = x.
\]

(27.7)

The derivative,

\[
D_2 y(t,x) = D_2 y(t,x)
\]

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 \( \rho_0(x) \). The following picture is useful to illustrate the ideas.
**Lemma 27.2.1** \( \rho_0(x) = \rho(t,y(t,x)) \det(F) \) 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 Lemma 27.7. Therefore, \( \det(F) = 1 \) at \( t = 0 \) and if it is assumed \( t \rightarrow \det(F) \) 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,

\[
\text{mass of } V_t = \int_{V_t} \rho(t,y) \ v(t,y) \ dV = \int_{V_0} \rho(t,y(t,x)) \det(F) \ dV
\]

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, \( V_t \) at time \( t \) which corresponds to \( V_0 \) at time \( t = 0 \). The total linear momentum of this material at time \( t \) is

\[
\int_{V_t} \rho(t,y) \ v(t,y) \ dV
\]

where \( v \) 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,
27.2. SOME APPLICATIONS OF THE DIVERGENCE THEOREM

By Lemma 27.2.41 and the change of variables formula for multiple integrals
\[
\frac{d}{dt} \left( \int_{V_t} \rho(t, y) \mathbf{v}(t, y) \, dV(y) \right) = \frac{d}{dt} \left( \int_{V_0} \rho(t, y(t, x)) \mathbf{v}(t, y(t, x)) \det(F) \, dV(x) \right)
\]
\[
= \frac{d}{dt} \left( \int_{V_0} \rho_0(x) \mathbf{v}(t, y(t, x)) \, dV(x) \right)
\]
\[
= \int_{V_0} \rho_0(x) \left[ \frac{\partial \mathbf{v}}{\partial t} + \nabla \cdot \mathbf{v} \right] \, dV(x)
\]
\[
= \int_{V_0} \rho_0(x) \left[ \frac{\partial \mathbf{v}}{\partial t} + \nabla \cdot \mathbf{v} \right] \, dV(x)
\]
\[
= \int_{V_0} \rho_0(x) \det(F) \left[ \frac{\partial \mathbf{v}}{\partial t} + \nabla \cdot \mathbf{v} \right] \, dV(x)
\]
\[
= \int_{V_0} \rho_0(x) \det(F) \left[ \frac{\partial \mathbf{v}}{\partial t} + \nabla \cdot \mathbf{v} \right] \, dV(x)
\]

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 acts on the boundary of the chunk of material called a traction force. Typically, the body force is something like gravity in which case, \( \mathbf{b} = -g \rho \mathbf{k} \), assuming the Cartesian coordinate system has been chosen in the usual manner. The traction force is of the form
\[
\int_{\partial V_t} \mathbf{s}(t, y, \mathbf{n}) \, dA
\]
where \( \mathbf{n} \) is the unit exterior normal. Thus the traction force depends on position, time, and the orientation of the boundary of \( V_t \). 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, \( T_{ij} = T_{ji} \). 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,
\[
\int_{V_t} \rho(t, y) \left[ \frac{\partial \mathbf{v}}{\partial t} + \nabla \cdot \mathbf{v} \right] \, dV(y) = \int_{V_t} \mathbf{b}(t, y) \, dV(y) + \int_{\partial V_t} T_{ij}(t, y) n_j \, dA.
\]
Here is where the divergence theorem is used. In the last integral, the multiplication by \( n_j \) is exchanged for the \( j^{th} \) partial derivative and an integral over \( V_t \). Thus
\[
\int_{V_t} \rho(t, y) \left[ \frac{\partial \mathbf{v}}{\partial t} + \nabla \cdot \mathbf{v} \right] \, dV(y) = \int_{V_t} \mathbf{b}(t, y) \, dV(y) + \int_{V_t} \frac{\partial (T_{ij}(t, y))}{\partial y_j} \, dV(y).
\]
Since \( V_t \) was arbitrary, it follows
\[
\rho(t, y) \left[ \frac{\partial \mathbf{v}}{\partial t} + \nabla \cdot \mathbf{v} \right] = \mathbf{b}(t, y) + \frac{\partial (T_{ij}(t, y))}{\partial y_j}
\]
\[
= \mathbf{b}(t, y) + \nabla \cdot T.
\]
where here \( \nabla \cdot T \) is a vector whose \( i^{th} \) component is given by
\[
(\nabla \cdot T)_i = \frac{\partial T_{ij}}{\partial y_j}.
\]
The term, \( \frac{\partial \mathbf{v}}{\partial t} + \nabla \cdot \mathbf{v} \), is the total derivative with respect to \( t \) of the velocity \( \mathbf{v} \). Thus you might see this written as
\[
\rho \dot{\mathbf{v}} = \mathbf{b} + \nabla \cdot T.
\]
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

\[
\int_{V_0} \rho_0 (x) v (t, x) \, dV
\]

and so, since \( x \) does not depend on \( t \),

\[
\frac{d}{dt} \left( \int_{V_0} \rho_0 (x) v (t, x) \, dV \right) = \int_{V_0} \rho_0 (x) v_t (t, x) \, dV.
\]

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

\[
\int_{V_0} \rho_0 (x) v_t (t, x) \, dV = \int_{V_0} b_0 (t, x) \, dV + \int_{\partial V_t} T_{ij} n_j dA,
\]

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 \( \partial V_0 \). For \( y \in \partial V_t \) there is a unit outer normal, \( n \). Here \( y = y (t, x) \) for \( x \in \partial V_0 \). Then define \( N \) to be the unit outer normal to \( V_0 \) at the point, \( x \). Near the point \( y \in \partial V_t \) the surface, \( \partial V_t \) 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 \( \partial V_t \) near this point is

\[
n = \frac{y_s (s,t) \times y_t (s,t)}{|y_s (s,t) \times y_t (s,t)|}
\]

with the area element given by \(|y_s (s,t) \times y_t (s,t)| \, ds \, dt\). This is true for \( y \in P_t \subseteq \partial V_t \), a small piece of \( \partial V_t \). Therefore, the last integral in (27.8) is the sum of integrals over small pieces of the form

\[
\int_{P_t} T_{ij} n_j dA
\]

where \( P_t \) is parametrized by \( y (s,t) \), \( (s,t) \in D \). Thus the integral in (27.8) is of the form

\[
\int_D T_{ij} (y (s,t)) \left( y_s (s,t) \times y_t (s,t) \right)_j \, ds \, dt.
\]

By the chain rule this equals

\[
\int_D T_{ij} (y (s,t)) \left( \frac{\partial y}{\partial x_\alpha} \frac{\partial x_\alpha}{\partial s} \times \frac{\partial y}{\partial x_\beta} \frac{\partial x_\beta}{\partial t} \right)_j \, ds \, dt.
\]

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 \( \partial V_t \) 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

\[
\int_D T_{ij} (x (s,t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \left( \frac{\partial y}{\partial x_\alpha} \times \frac{\partial y}{\partial x_\beta} \right)_j \, ds \, dt
\]

where here \( x (s,t) \) is the point of \( \partial V_0 \) which corresponds with \( y (s,t) \in \partial V_t \). Thus

\[
T_{ij} (x (s,t)) = T_{ij} (y (s,t)).
\]
27.2. SOME APPLICATIONS OF THE DIVERGENCE THEOREM

(Perhaps this is a slight abuse of notation because \( T_{ij} \) is defined on \( \partial V_t \), not on \( \partial V_0 \), but it avoids introducing extra symbols.) Next we have

\[
\int_D T_{ij} (x(s,t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \varepsilon_{pab} \frac{\partial y_a}{\partial x_\alpha} \frac{\partial y_b}{\partial x_\beta} \, ds \, dt
\]

\[
= \int_D T_{ij} (x(s,t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \varepsilon_{cab} \delta_{jc} \frac{\partial y_a}{\partial x_\alpha} \frac{\partial y_b}{\partial x_\beta} \, ds \, dt
\]

\[
= \int_D T_{ij} (x(s,t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \varepsilon_{cab} \delta_{jc} \frac{\partial y_a}{\partial x_\alpha} \frac{\partial y_b}{\partial x_\beta} \, ds \, dt
\]

\[
= \int_D T_{ij} (x(s,t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \varepsilon_{pab} \det(F) \, ds \, dt.
\]

Now \( \frac{\partial x_\alpha}{\partial y_j} = F^{-1}_{pj} \) and also

\[
\varepsilon_{pab} \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} = (x_s \times x_t)_p
\]

so the result just obtained is of the form

\[
\int_D (\det F) F^{-1}_{pj} T_{ij} (x(s,t)) (x_s \times x_t)_p \, ds \, dt =
\]

\[
\int_D (\det F) T_{ij} (x(s,t)) (F^{-T})_{jp} (x_s \times x_t)_p \, ds \, dt.
\]

This has transformed the integral over \( P_t \) to one over \( P_0 \), the part of \( \partial V_0 \) which corresponds with \( P_t \). Thus the last integral is of the form

\[
\int_{P_0} (\det F) (T F^{-T})_{ip} N_p dA
\]

Summing these up over the pieces of \( \partial V_t \) and \( \partial V_0 \) yields the last integral in \( \mathbb{E} \) equals

\[
\int_{\partial V_0} (\det F) (T F^{-T})_{ip} N_p dA
\]

and so the balance of momentum in terms of the material coordinates becomes

\[
\int_{V_0} \rho_0 (x) v_t (t, x) \, dV = \int_{V_0} b_0 (t, x) \, dV + \int_{\partial V_0} (\det F) (T F^{-T})_{ip} N_p dA
\]

The matrix, \( (\det F) (T F^{-T})_{ip} \) is called the Piola Kirchhoff stress, \( S \). An application of the divergence theorem yields

\[
\int_{V_0} \rho_0 (x) v_t (t, x) \, dV = \int_{V_0} b_0 (t, x) \, dV + \int_{V_0} \frac{\partial \left( (\det F) (T F^{-T})_{ip} \right)}{\partial x_p} \, dV.
\]
Since $V_0$ is arbitrary, a balance law for momentum in terms of the material coordinates is obtained
\[ \rho_0 (x) v_t (t, x) = b_0 (t, x) + \frac{\partial}{\partial x_p} \left( \det (F) \left( TF^{-T} \right)_{ip} \right) \]
\[ = b_0 (t, x) + \text{div} \left( \det (F) \left( TF^{-T} \right) \right) \]
\[ = b_0 (t, x) + \text{div} S. \tag{27.11} \]

As just shown, the relation between the Cauchy stress and the Piola Kirchhoff stress is
\[ S = \det (F) \left( TF^{-T} \right), \tag{27.12} \]
perhaps not the first thing you would think of.

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 are known as constitutive laws.

### 27.2.6 Frame Indifference

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 for the spatial coordinates was chosen.

For $Q(t)$ an orthogonal transformation and
\[ y' = q(t) + Q(t) y, \quad n' = Qn, \]
the new spatial coordinates are denoted by $y'$. Recall an orthogonal transformation is just one which satisfies
\[ Q(t)^T Q(t) = Q(t) Q(t)^T = I. \]
The stress has to do with the traction force area density produced by internal changes in the body and has nothing to do with the way the body is observed. Therefore, it is required that
\[ T' n' = QTn. \]

Thus
\[ T' Qn = QTn. \]

Since this is true for any $n$ normal to the boundary of any piece of the material considered, it must be the case that
\[ T' Q = QT \]
and so
\[ T' = Q T Q^T. \]

This is called frame indifference.

By (27.12), the Piola Kirchhoff stress, $S$ is related to $T$ by
\[ S = \det (F) TF^{-T}, \quad F = D_y x. \]

This stress involves the use of the material coordinates and a normal $N$ to a piece of the body in reference configuration. Thus $S N$ gives the force on a part of $\partial V_t$ per unit area on $\partial V_0$. Then for a different choice of spatial coordinates, $y' = q(t) + Q(t) y$,
\[ S' = \det (F') T' (F')^{-T}. \]
27.2. SOME APPLICATIONS OF THE DIVERGENCE THEOREM

but

\[ F' = D_x Y' = Q(t) D_x Y = QF \]

and so frame indifference in terms of \( S \) is

\[
S' = \det (F) QTQ^T (QF)^{-T} \\
= \det (F) QTQ^T QF^{-T} \\
= QS
\]

This principle of frame indifference is sometimes ignored and there are certainly interesting mathematical models which have resulted from doing this, but such things cannot be considered physically acceptable.

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.

27.2.7 Bernoulli’s Principle

Consider a possibly moving fluid with constant density, \( \rho \) and let \( P \) denote the pressure in this fluid. If \( B \) is a part of this fluid the force exerted on \( B \) by the rest of the fluid is \( \int_{\partial B} -Pn \, dA \) where \( n \) is the outer normal from \( B \). Assume this is the only force which matters so for example there is no viscosity in the fluid. Thus the Cauchy stress in rectangular coordinates should be

\[
T = \begin{pmatrix} -P & 0 & 0 \\ 0 & -P & 0 \\ 0 & 0 & -P \end{pmatrix}.
\]

Then

\[ \text{div} \, T = -\nabla P. \]

Also suppose the only body force is from gravity, a force of the form

\[ -\rho g k \]

and so from the balance of momentum

\[ \rho \dot{v} = -\rho g k - \nabla P(x). \quad (27.13) \]

Now in all this the coordinates are the spatial coordinates and it is assumed they are rectangular. Thus

\[ x = (x, y, z)^T \]

and \( \mathbf{v} \) is the velocity while \( \dot{\mathbf{v}} \) is the total derivative of \( \mathbf{v} = (v_1, v_2, v_3)^T \) given by \( \mathbf{v}_t + v_i v_{,i} \). Take the dot product of both sides of \( \mathbf{v} \) with \( \mathbf{v} \). This yields

\[
(\rho/2) \frac{d}{dt} |\mathbf{v}|^2 = -\rho g \frac{dz}{dt} - \frac{d}{dt} P(x).
\]

Therefore,

\[
\frac{d}{dt} \left( \frac{\rho |\mathbf{v}|^2}{2} + \rho g z + P(x) \right) = 0
\]
and so there is a constant, \( C' \) such that
\[
\frac{\rho |v|^2}{2} + \rho g z + P(x) = C'
\]
For convenience define \( \gamma \) to be the weight density of this fluid. Thus \( \gamma = \rho g \). Divide by \( \gamma \). Then
\[
\frac{|v|^2}{2g} + \frac{z}{\gamma} + \frac{P(x)}{\gamma} = C.
\]
this is Bernoulli's\(^3\) principle. Note how if you keep the height the same, then if you raise \(|v|\), it follows the pressure drops.

This is often used to explain the lift of an airplane wing. The top surface is curved which forces the air to go faster over the top of the wing causing a drop in pressure which creates lift. It is also used to explain the concept of a venturi tube in which the air loses pressure due to being pinched which causes it to flow faster. In many of these applications, the assumptions used in which \( \rho \) is constant and there is no other contribution to the traction force on \( \partial B \) than pressure so in particular, there is no viscosity, are not correct. However, it is hoped that the effects of these deviations from the ideal situation above are small enough that the conclusions are still roughly true. You can see how using balance of momentum can be used to consider more difficult situations. For example, you might have a body force which is more involved than gravity.

### 27.2.8 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 \( u \equiv y - x \). Thus \( u \) is the displacement. Then from (27.11) you can verify the following holds.
\[
\rho_0(x) u_{tt}(t, x) = b_0(t, x) + k \Delta u(t, x) \tag{27.14}
\]
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.

### 27.2.9 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. 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. The possibility exists that the

\(^3\)There were many Bernoullis. This is Daniel Bernoulli. He seems to have been nicer than some of the others. Daniel was actually a doctor who was interested in mathematics. He lived from 1700-1782.
27.2. SOME APPLICATIONS OF THE DIVERGENCE THEOREM*

reason for this is the numerical methods used to solve the partial differential equations may be better physical models than the balance laws themselves. It is an area where people still sometimes disagree.

27.2.10 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,

$$E = \nabla \left( qk \frac{1}{|x - x_0|} \right).$$

The other thing which is significant about this is that in three dimensions and for $x \neq x_0$,

$$\nabla \cdot \nabla \left( qk \frac{1}{|x - x_0|} \right) = \nabla \cdot E = 0. \quad (27.15)$$

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,

$$\int_{\partial B_\varepsilon} \mathbf{E} \cdot \mathbf{n} dS = -\int_{\partial B_\varepsilon} k \frac{q (x-x_0) \cdot \frac{x-x_0}{|x-x_0|}}{|x-x_0|^3} dS$$

$$= -kq \int_{\partial B_\varepsilon} \frac{1}{|x-x_0|^2} dS = \frac{-kq}{\varepsilon^2} \int_{\partial B_\varepsilon} dS$$

$$= \frac{-kq}{\varepsilon^2} 4\pi \varepsilon^2 = -4\pi k.$$ 

Therefore, from the divergence theorem and observation 27.15,

$$-4\pi kq + \int_{\partial U} \mathbf{E} \cdot \mathbf{n} dS = \int_{\partial U_\varepsilon} \mathbf{E} \cdot \mathbf{n} dS = \int_{U_\varepsilon} \nabla \cdot \mathbf{E} dV = 0.$$ 

It follows that

$$4\pi kq = \int_{\partial U} \mathbf{E} \cdot \mathbf{n} dS.$$ 

If there are several charges located inside $U$, say $q_1, q_2, \ldots, q_n$, then letting $\mathbf{E}_i$ 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,

$$\int_{\partial U} \mathbf{E} \cdot \mathbf{n} dS = \sum_{i=1}^{n} \int_{\partial U} \mathbf{E}_i \cdot \mathbf{n} dS$$

$$= \sum_{i=1}^{n} 4\pi k q_i = 4\pi k \sum_{i=1}^{n} q_i.$$ 

This is known as Gauss’s law and it is the fundamental result in electrostatics.
Chapter 28

Stokes And Green’s Theorems

28.1 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 seemingly more general Stokke’s theorem but is interesting for it’s own sake. Historically, theorems like it were 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 28.1.1** 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

\[
\int_{\partial U} P \, dx + Q \, dy = \int_U \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) \, dA. \tag{28.1}
\]

**Proof:** First consider the right side of (28.1).

\[
\int_U \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) \, dA
\]
\[= \int_c^{r(y)} \partial Q \frac{\partial}{\partial x} \, dx \, dy - \int_a^{b(x)} \partial P \frac{\partial}{\partial y} \, dy \, dx\]
\[= \int_c^d (Q(r(y), y) - Q(l(y), y)) \, dy + \int_a^b (P(x, b(x))) - P(x, t(x)) \, dx. \quad (28.2)\]

Now consider the left side of (28.1). Denote by \(V\) the vertical parts of \(\partial U\) and by \(H\) the horizontal parts.

\[\int_{\partial U} F \cdot d\mathbf{R} = \]
\[= \int_c^d (0, Q(r(s), s)) \cdot (r'(s), 1) \, ds + \int_H (0, Q(r(s), s)) \cdot (\pm1, 0) \, ds\]
\[- \int_c^d (0, Q(l(s), s)) \cdot (l'(s), 1) \, ds + \int_a^b (P(s, b(s)), 0) \cdot (1, b'(s)) \, ds\]
\[+ \int_V (P(s, b(s)), 0) \cdot (0, \pm1) \, ds - \int_a^b (P(s, t(s)), 0) \cdot (1, t'(s)) \, ds\]
\[= \int_c^d Q(r(s), s) \, ds - \int_c^d Q(l(s), s) \, ds + \int_a^b P(s, b(s)) \, ds - \int_a^b P(s, t(s)) \, ds\]

which coincides with (28.2). This proves the lemma.

**Corollary 28.1.2** Let everything be the same as in Lemma 24.0.9 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 (28.1) is valid for any triangle for example.

Now suppose (28.1) holds for \(U_1, U_2, \cdots, U_m\) 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 (28.1) must hold for \(U \equiv \bigcup_{i=1}^m U_i\), the union of these sets. This is because

\[\int_U \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) \, dA = \]
\[= \sum_{k=1}^m \int_{U_k} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) \, dA\]
\[= \sum_{k=1}^m \int_{\partial U_k} \mathbf{F} \cdot d\mathbf{R} = \int_{\partial U} \mathbf{F} \cdot d\mathbf{R}\]

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 28.1 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 28.1. Roughly speaking, you can drill holes in a region for which 28.1 holds and get another region for which this continues to hold provided 28.1 holds for the holes. To see why this is so, consider the following picture which typifies the situation just described.

Then
\[
\int_{\partial V} \mathbf{F} \cdot d\mathbf{R} = \int_V \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA
\]
\[
= \int_U \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA + \int_{V \setminus U} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA
\]
\[
= \int_{\partial U} \mathbf{F} \cdot d\mathbf{R} + \int_{V \setminus U} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA
\]
and so
\[
\int_{V \setminus U} \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \int_{\partial V} \mathbf{F} \cdot d\mathbf{R} - \int_{\partial U} \mathbf{F} \cdot d\mathbf{R}
\]
which equals
\[
\int_{\partial (V \setminus U)} \mathbf{F} \cdot d\mathbf{R}
\]
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 28.1 is valid quite generally. This verifies the following theorem.

**Theorem 28.1.3 (Green’s Theorem)** Let $U$ be an open set in the plane and let $\partial U$ be piecewise smooth and let $\mathbf{F}(x,y) = (P(x,y),Q(x,y))$ be a $C^1$ vector field defined near $U$. Then it is often\footnote{For a general version see the advanced calculus book by Apostol. The general versions involve the concept of a rectifiable (finite length) Jordan curve.} the case that
\[
\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} = \int_U \left( \frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y) \right) dA.
\]
Here is an alternate proof of Green’s theorem from the divergence theorem.

**Theorem 28.1.4 (Green’s Theorem)** Let \( U \) be an open set in the plane and let \( \partial U \) be piecewise smooth and let \( \mathbf{F}(x,y) = (P(x,y),Q(x,y)) \) be a \( C^1 \) vector field defined near \( U \). Then it is often the case that

\[
\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} = \int_U \left( \frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y) \right) dA.
\]

**Proof:** Suppose the divergence theorem holds for \( U \). Consider the following picture.

Since it is assumed that motion around \( U \) is counter clockwise, the tangent vector, \((x',y')\) is as shown. The unit **exterior normal** is a multiple of

\[
(x', y', 0) \times (0, 0, 1) = (y', -x', 0).
\]

Use your right hand and the geometric description of the cross product to verify this. This would be the case at all the points where the unit exterior normal exists.

Now let \( \mathbf{F}(x,y) = (Q(x,y),-P(x,y)) \). Also note the area (length) element on the bounding curve \( \partial U \) is \( \sqrt{(x')^2 + (y')^2} \, dt \). Suppose the boundary of \( U \) consists of \( m \) smooth curves, the \( i^{th} \) of which is parametrized by \((x_i, y_i)\) with the parameter, \( t \in [a_i, b_i] \). Then by the divergence theorem,

\[
\int_U (Q_x - P_y) \, dA = \int_U \text{div} \,(\mathbf{F}) \, dA = \int_{\partial U} \mathbf{F} \cdot \mathbf{n} \, dS
\]

\[
= \sum_{i=1}^{m} \int_{a_i}^{b_i} \left( Q(x_i(t), y_i(t)) - P(x_i(t), y_i(t)) \right) \frac{1}{\sqrt{(x_i')^2 + (y_i')^2}} \sqrt{(x_i')^2 + (y_i')^2} \, dt
\]

\[
= \sum_{i=1}^{m} \int_{a_i}^{b_i} \left( Q(x_i(t), y_i(t)) - P(x_i(t), y_i(t)) \right) \cdot (y_i', -x_i') \, dt
\]

\[
= \sum_{i=1}^{m} \int_{a_i}^{b_i} Q(x_i(t), y_i(t)) y_i'(t) + P(x_i(t), y_i(t)) x_i'(t) \, dt = \int_{\partial U} P \, dx + Q \, dy
\]

This proves Green’s theorem from the divergence theorem.
**Proposition 28.1.5** 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 28.1.6** Use Proposition 24.2.1 to find the area of the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1.$$

You can parametrize the boundary of this ellipse as

$$x = a \cos t, \quad y = b \sin t, \quad t \in [0, 2\pi].$$

Then from Proposition 24.2.1,

\[
\text{Area equals} = \frac{1}{2} \int_{0}^{2\pi} (-b \sin t, a \cos t) \cdot (-a \sin t, b \cos t) \, dt = \frac{1}{2} \int_{0}^{2\pi} (a b) \, dt = \pi a b.
\]

**Example 28.1.7** Find $\int_{\partial U} F \cdot dR$ where $U$ is the set,

$$\{(x, y) : x^2 + 3y^2 \leq 9\}$$

and $F(x, y) = (y, -x)$.

One way to do this is to parametrize the boundary of $U$ and then compute the line integral directly. It is easier to use Green’s theorem. The desired line integral equals

$$\int_{U} ((-1) - 1) \, dA = -2 \int_{U} dA.$$ 

Now $U$ is an ellipse having area equal to $3\sqrt{3}$ and so the answer is $-6\sqrt{3}$.

**Example 28.1.8** Find $\int_{\partial U} F \cdot dR$ where $U$ is the set, $\{(x, y) : 2 \leq x \leq 4, 0 \leq y \leq 3\}$ and $F(x, y) = (x \sin y, y^3 \cos x)$.

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 28.1.9** Find $\int_{\partial U} F \cdot dR$ where $U$ is the set,

$$\{(x, y) : 2 \leq x \leq 4, x \leq y \leq 3\}$$

and $F(x, y) = (x \sin y, y \sin x)$.

From Green’s theorem this line integral equals

\[
\int_{2}^{4} \int_{x}^{3} (y \cos x - x \cos y) \, dy \, dx = -\frac{3}{2} \sin 4 - 6 \sin 3 - 8 \cos 4 - \frac{9}{2} \sin 2 + 4 \cos 2.
\]
28.2 Exercises

1. Find \( \int_S x\,dS \) where \( S \) is the surface which results from the intersection of the cone \( z = 2 - \sqrt{x^2 + y^2} \) with the cylinder \( x^2 + y^2 - 2x = 0 \).

2. Now let \( n \) be the unit normal to the above surface which has positive \( z \) component and let \( F(x, y, z) = (x, y, z) \). Find the flux integral,

\[
\int_S F \cdot ndS.
\]

3. Find \( \int_S z\,dS \) where \( S \) is the surface which results from the intersection of the hemisphere \( z = \sqrt{4 - x^2 - y^2} \) with the cylinder \( x^2 + y^2 - 2x = 0 \).

4. In the situation of the above problem, find the flux integral

\[
\int_S F \cdot ndS
\]

where \( n \) is the unit normal to the surface which has positive \( z \) component and \( F = (x, y, z) \).

5. Let \( x^2/a^2 + y^2/b^2 = 1 \) be an ellipse. Show using Green’s theorem that its area is \( \pi ab \).

6. A spherical storage tank having radius \( a \) is filled with water which weights 62.5 pounds per cubic foot. It is shown later that this implies that the pressure of the water at depth \( z \) equals 62.5z. Find the total force acting on this storage tank.

7. Let \( n \) be the unit normal to the cone \( z = \sqrt{x^2 + y^2} \) which has negative \( z \) component and let \( F = (x, 0, z) \) be a vector field. Let \( S \) be the part of this cone which lies between the planes \( z = 1 \) and \( z = 2 \).
Find
\[ \int_S \mathbf{F} \cdot \mathbf{n} dS \]

8. Let \( S \) be the surface \( z = 9 - x^2 - y^2 \) for \( x^2 + y^2 \leq 9 \). Let \( \mathbf{n} \) be the unit normal to \( S \) which points up. Let \( \mathbf{F} = (y, -x, z) \) and find
\[ \int_S \mathbf{F} \cdot \mathbf{n} dS \]

9. Let \( S \) be the surface \( 3z = 9 - x^2 - y^2 \) for \( x^2 + y^2 \leq 9 \). Let \( \mathbf{n} \) be the unit normal to \( S \) which points up. Let \( \mathbf{F} = (y, -x, z) \) and find
\[ \int_S \mathbf{F} \cdot \mathbf{n} dS \]

10. For \( \mathbf{F} = (x, y, z) \), \( S \) is the part of the cylinder \( x^2 + y^2 = 1 \) between the planes \( z = 1 \) and \( z = 3 \). Letting \( \mathbf{n} \) be the unit normal which points away from the \( z \) axis, find
\[ \int_S \mathbf{F} \cdot \mathbf{n} dS \]

11. Let \( S \) be the part of the sphere of radius \( a \) which lies between the two cones \( \phi = \frac{\pi}{4} \) and \( \phi = \frac{\pi}{6} \). Let \( \mathbf{F} = (z, y, 0) \). Find the flux integral \( \int_S \mathbf{F} \cdot \mathbf{n} dS \).

12. Let \( S \) be the part of a sphere of radius \( a \) above the plane \( z = \frac{a}{2} \), \( \mathbf{F} = (2x, 1, 1) \) and let \( \mathbf{n} \) be the unit upward normal on \( S \). Find \( \int_S \mathbf{F} \cdot \mathbf{n} dS \).

13. In the above, problem, let \( C \) be the boundary of \( S \) oriented counter clockwise as viewed from high on the \( z \) axis. Find
\[ \int_C 2xdx + dy + dz \]

14. Let \( S \) be the top half of a sphere of radius \( a \) centered at \( \mathbf{0} \) and let \( \mathbf{n} \) be the unit outward normal. Let \( \mathbf{F} = (0, 0, z) \). Find
\[ \int_S \mathbf{F} \cdot \mathbf{n} dS \]

15. Let \( D \) be a circle in the plane which has radius 1 and let \( C \) be its counter clockwise boundary. Find
\[ \int_C ydx + xdy \]

16. Let \( D \) be a circle in the plane which has radius 1 and let \( C \) be its counter clockwise boundary. Find
\[ \int_C ydx - xdy \]

17. Find
\[ \int_C (x + y) dx \]

where \( C \) is the square curve which goes from \((0, 0) \to (1, 0) \to (1, 1) \to (0, 1) \to (0, 0) \).
18. Find the line integral $\int_C (\sin x + y) \, dx + y^2 \, dy$ where $C$ is the oriented square

$$(0, 0) \to (1, 0) \to (1, 1) \to (0, 1) \to (0, 0).$$

19. Let $P(x, y) = \frac{-y}{x^2+y^2}, Q(x, y) = \frac{x}{x^2+y^2}$. Show $Q_x - P_y = 0$. Let $D$ be the unit disk. Compute directly $\int_C P \, dx + Q \, dy$ where $C$ is the counter clockwise circle of radius 1 which bounds the unit disk. Why don’t you get 0 for the line integral?

20. Let $F = (2y, \ln (1 + y^2) + x)$. Find $\int_C F \cdot dR$ where $C$ is the curve consisting of line segments,

$$(0, 0) \to (1, 0) \to (1, 1) \to (0, 0).$$

28.3 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 28.3.1** 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

$$\nabla \times F = \begin{vmatrix} i & j & k \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{vmatrix} = (\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}) i + (\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}) j + (\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}) k.$$

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 28.3.2** 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 $V$.

$$(R_u \times R_v) \cdot (\nabla \times F)(R(u, v)) = ((F \circ R)_u \cdot R_v - (F \circ R)_v \cdot R_u)(u, v). \quad (28.3)$$
28.3. STOKE’S THEOREM FROM GREEN’S THEOREM

Proof: Start with the left side and let \( x_i = R_i(u,v) \) for short.

\[
(R_u \times R_v) \cdot (\nabla \times F)(R(u,v)) = \epsilon_{ijk} x_{ju} x_{kv} \epsilon_{irs} \frac{\partial F_s}{\partial x_r}
\]

\[
= (\delta_{jr} \delta_{ks} - \delta_{js} \delta_{kr}) x_{ju} x_{kv} \frac{\partial F_r}{\partial x_r}
\]

\[
= x_{ju} x_{kv} \frac{\partial F_k}{\partial x_j} - x_{ju} x_{kv} \frac{\partial F_j}{\partial x_k}
\]

\[
= R_v \cdot \frac{\partial (F \circ R)}{\partial u} - R_u \cdot \frac{\partial (F \circ R)}{\partial v}
\]

which proves (28.3).

The proof of Stoke’s theorem given next follows [4]. First, it is convenient to give a definition.

**Definition 28.3.3** A vector valued function, \( R : U \subseteq \mathbb{R}^m \rightarrow \mathbb{R}^n \) 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 \( k \).

Let \( U \) be a region in \( \mathbb{R}^2 \) for which Green’s theorem holds. Thus Green’s theorem says that if \( P, Q \) are in \( C^1(U) \),

\[
\int_U (Q_u - P_v) \, dA = \int_{\partial U} f \cdot dR
\]

Here the \( u \) and \( v \) axes are in the same relation as the \( x \) and \( y \) axes. That is, the following picture holds. That is, the positive \( x \) and \( u \) axes both point to the right and the positive \( y \) and \( v \) axes point up.

![Diagram showing the axes relation](image)

**Theorem 28.3.4** (Stoke’s Theorem) Let \( U \) be any region in \( \mathbb{R}^2 \) for which the conclusion of Green’s theorem holds and let \( R \in C^2(\overline{U}, \mathbb{R}^3) \) be a one to one function satisfying \( |(R_u \times R_v)(u,v)| \neq 0 \) for all \( (u,v) \in U \) and let \( S \) denote the surface,

\[
S = \{ R(u,v) : (u,v) \in U \},
\]

\[
\partial S = \{ R(u,v) : (u,v) \in \partial U \}
\]

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 \( \partial U \) as described above). Then for \( F \) a \( C^1 \) vector field defined near \( S \),

\[
\int_{\partial S} F \cdot dR = \int_S \text{curl}(F) \cdot n \, dS
\]

where \( n \) is the normal to \( S \) defined by

\[
n = \frac{R_u \times R_v}{|R_u \times R_v|}.
\]
**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 $C$,

$$ \int_{R(C)} F \cdot dR = $$

$$ = \int_{\alpha}^{\beta} F(R(u(t), v(t))) \cdot (R_u u'(t) + R_v v'(t)) \, dt $$

$$ = \int_{\alpha}^{\beta} F(R(u(t), v(t))) R_u (u(t), v(t)) u'(t) \, dt $$

$$ + \int_{\alpha}^{\beta} F(R(u(t), v(t))) R_v (u(t), v(t)) v'(t) \, dt $$

$$ = \int_{C} ((F \circ R) \cdot R_u, (F \circ R) \cdot R_v) \cdot dr. $$

Since this holds for each such piece of $\partial U$, it follows

$$ \int_{\partial S} F \cdot dR = \int_{\partial U} ((F \circ R) \cdot R_u, (F \circ R) \cdot R_v) \cdot dr. $$

By the assumption that the conclusion of Green’s theorem holds for $U$, this equals

$$ \int_{U} [(F \circ (R_u) \cdot R_v)] u A $$

$$ = \int_{U} [((F \circ R) \cdot R_v) u A - (F \circ R) \cdot R_{uv} - (F \circ R) \cdot R_{vu} - (F \circ R) \cdot R_{uv} - (F \circ R) \cdot R_{uv}] A $$

$$ = \int_{U} [((F \circ R) \cdot R_v) u A - (F \circ R) \cdot R_{uv} - (F \circ R) \cdot R_{uv}] A $$

the last step holding by equality of mixed partial derivatives, a result of the assumption that $R$ is $C^2$. Now by Lemma 28.3.2, this equals

$$ \int_{U} (R_u \times R_v) \cdot (\nabla \times F) \, dA $$

$$ = \int_{U} \nabla \times F \cdot (R_u \times R_v) \, dA $$

$$ = \int_{S} \nabla \times F \cdot n \, dS $$

because $dS = |(R_u \times R_v)| \, dA$ and $n = \frac{(R_u \times R_v)}{|(R_u \times R_v)|}$. Thus

$$ (R_u \times R_v) \, dA = \frac{(R_u \times R_v)}{|(R_u \times R_v)|} |(R_u \times R_v)| \, dA $$

$$ = n \, dS. $$

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.
28.3.1 The Normal And The Orientation

Stoke’s theorem as just presented needs no apology. However, it is helpful in applications to have some additional geometric insight.

To begin with, suppose the surface $S$ of interest is a parallelogram in $\mathbb{R}^3$ determined by the two vectors $a, b$. Thus $S = R(u,v)$ where $Q = [0,1] \times [0,1]$ is the unit square and for $(u,v) \in Q$,

$$R(u,v) \equiv ua + vb + p,$$

the point $p$ being a corner of the parallelogram $S$. Then orient $\partial S$ consistent with the counter clockwise orientation on $\partial Q$. Thus, following this orientation on $S$ you go from $p$ to $p + a$ to $p + a + b$ to $p + b$ to $p$. Then Stoke’s theorem implies that with this orientation on $\partial S$,

$$\int_{\partial S} F \cdot dR = \int_S \nabla \times F \cdot n ds$$

where

$$n = R_u \times R_v / |R_u \times R_v| = a \times b / |a \times b|.$$

Now recall $a, b, a \times b$ forms a right hand system.

Thus, if you were walking around $\partial S$ in the direction of the orientation with your left hand over the surface $S$, the normal vector $a \times b$ would be pointing in the direction of your head.

More generally, if $S$ is a surface which is not necessarily a parallelogram but is instead as described in Theorem 28.3.4, you could consider a small rectangle $Q$ contained in $U$ and orient the boundary of $R(Q)$ consistent with the counter clockwise orientation on $\partial Q$. Then if $Q$ is small enough, as you walk around $\partial R(Q)$ in the direction of the described orientation with your left hand over $R(Q)$, your head points roughly in the direction of $R_u \times R_v$.

As explained above, this is true of the tangent parallelogram, and by continuity of $R_u, R_v$, the normals to the surface $R(Q) R_u \times R_v (u)$ for $u \in Q$ will still point roughly in the same direction as your head if you walk in the indicated direction over $\partial R(Q)$, meaning the angle between the vector from your feet to your head and the vector $R_u \times R_v (u)$ is less than $\pi/2$.

You can imagine filling $U$ with such non-overlapping regions $Q_i$. Then orienting $\partial R(Q_i)$ consistent with the counter clockwise orientation on $Q_i$, and adding the resulting line integrals, the line integrals over the common sides cancel as indicated in the following picture and the result is the line integral over $\partial S$. 

---

### Diagram

[Diagram of a parallelogram and associated vectors]
Thus there is a simple relation between the field of normal vectors on $S$ and the orientation of $\partial S$. It is simply this. If you walk along $\partial S$ in the direction mandated by the orientation, with your left hand over the surface, the nearby normal vectors in Stoke’s theorem will point roughly in the direction of your head.

This also illustrates that you can define an orientation for $\partial S$ by specifying a field of unit normal vectors for the surface, which varies continuously over the surface, and require that the motion over the boundary of the surface is such that your head points roughly in the direction of nearby normal vectors as you walk along the boundary with your left hand over $S$. The existence of such a continuous field of normal vectors is what constitutes an orientable surface.

### 28.3.2 The Mobius Band

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 of unit normals to the surface which is a continuous function of position on $S$.

An example of a surface which is not orientable is the famous Mobius band, obtained by taking a long rectangular piece of paper and glueing the ends together after putting a twist in it. Here is a picture of one.

There is something quite interesting about this Mobius 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

\[ \mathbf{R}(\theta, v) \equiv \begin{cases} 
  x = 4 \cos \theta + v \cos \frac{\theta}{2} \\
  y = 4 \sin \theta + v \cos \frac{\theta}{2} \\
  z = v \sin \frac{\theta}{2}
\end{cases}, \quad \theta \in [0, 2\pi], \quad v \in [-1, 1]. \]

An obvious question is why the normal vector, \( \mathbf{R}_\theta \times \mathbf{R}_v / |\mathbf{R}_\theta \times \mathbf{R}_v| \) 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, \( \mathbf{R} \) is not one to one. In fact, \( \mathbf{R}(0, 0) = \mathbf{R}(2\pi, 0) \). Therefore, near this point on \( S \), there are two different values for the above normal vector. In fact, a tedious computation will show that this normal vector is

\[
\frac{(4 \sin \frac{1}{2} \theta \cos \theta - \frac{1}{2} v, 4 \sin \frac{1}{2} \theta \sin \theta + \frac{1}{2} v, -8 \cos^2 \frac{1}{2} \theta \sin \frac{1}{2} \theta - 8 \cos^3 \frac{1}{2} \theta + 4 \cos \frac{1}{2} \theta)}{D}
\]

where

\[
D = \left( 16 \sin^2 \left( \frac{\theta}{2} \right) + \frac{v^2}{2} + 4 \sin \left( \frac{\theta}{2} \right) v (\sin \theta - \cos \theta) \right.
\]
\[
+ 4^3 \cos^2 \left( \frac{\theta}{2} \right) \left( \cos \left( \frac{1}{2} \theta \right) \sin \left( \frac{1}{2} \theta \right) + \cos^2 \left( \frac{1}{2} \theta \right) - \frac{1}{2} \right)^2 \right)
\]

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 \( \mathbf{R} \) is one to one.

The Mobeus 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 \( \mathbf{R}_\theta \times \mathbf{R}_v / |\mathbf{R}_\theta \times \mathbf{R}_v| \) fails to deliver an orientation for the Mobeus 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 Mobeus 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 Ouyang when he was at meetings in Paris and saw it at a museum.

![Möbius strip](image1)

Here is an example of the use of Stoke’s theorem.

**Example 28.3.5** Let \( \mathbf{F} = (x + y, z, y) \). Let \( S \) be the top half of the ellipsoid

\[
\frac{x^2}{4} + \frac{y^2}{9} + z^2 = 1.
\]

Use Stoke’s theorem to evaluate the flux integral

\[
\int_S \text{curl} \,(\mathbf{F}) \cdot \mathbf{n} \, dS
\]
where \( \mathbf{n} \) is the unit normal to \( S \) which has positive \( z \) component. Also compute this directly.

First, a short computation shows that the curl is \( \begin{pmatrix} 0 & 0 & -1 \end{pmatrix}^T \). Now a parametrization of the surface is \( x = 2 \sin \phi \cos \theta, y = 3 \sin \phi \sin \theta, z = \cos \phi. \) \((\phi, \theta) \in [0, \frac{\pi}{2}] \times [0, 2\pi]\). Then

\[
\frac{\partial}{\partial \phi} \begin{pmatrix} 2 \sin \phi \cos \theta \\ 3 \sin \phi \sin \theta \\ \cos \phi \end{pmatrix} = \begin{pmatrix} 2 \cos \theta \cos \phi \\ 3 \cos \phi \sin \theta \\ -\sin \phi \end{pmatrix}
\]
\[
\frac{\partial}{\partial \theta} \begin{pmatrix} 2 \sin \phi \cos \theta \\ 3 \sin \phi \sin \theta \\ \cos \phi \end{pmatrix} = \begin{pmatrix} -2 \sin \theta \sin \phi \\ 2 \cos \theta \sin \phi \\ 3 \cos \theta \sin \phi \end{pmatrix}
\]

Then consider

\[
\begin{pmatrix} 2 \cos \theta \cos \phi \\ 3 \cos \phi \sin \theta \\ -\sin \phi \end{pmatrix} \times \begin{pmatrix} -2 \sin \theta \sin \phi \\ 2 \cos \theta \sin \phi \\ 3 \cos \theta \sin \phi \end{pmatrix} = \begin{pmatrix} 3 \cos \theta \sin^2 \phi \\ 6 \cos \phi \sin \phi \end{pmatrix}
\]

Is this the one we want? Yes, it has positive \( z \) component. Thus the integral to do is just

\[
\int_0^{\pi/2} \int_0^{2\pi} (6 \cos \phi \sin \phi) \, d\theta d\phi = -6\pi
\]

Now consider doing it using Stoke’s theorem. The curve bounding this surface is

\[
x^2 + \frac{y^2}{9} = 1
\]

Parametrizing this yields

\( x = 2 \cos t, y = 3 \sin t \)

for \( t \in [0, 2\pi] \). Then, since it goes in the correct direction to correspond with \( \mathbf{n} \) the field of normals which points up, it follows from Stokes theorem that the flux integral of the curl above is

\[
\int_C \mathbf{F} \cdot d\mathbf{r} = \int_0^{2\pi} (2 \cos t + 3 \sin t, 0, 3 \sin t) \cdot (-2 \sin t, 3 \cos t, 0) \, dt = -6\pi
\]

It could be argued that this was easier than the first way. However, there is an even easier way to think of this.

It is typically the case that \( \int_S \text{curl} \mathbf{F} \cdot \mathbf{n} \, dS = 0 \) whenever \( S \) is a closed surface. Just make a closed curve which goes around the surface. Then apply Stoke’s theorem on the top and on the bottom. This results in two line integrals around the closed curve which have opposite orientation. Hence their sum is 0 which equals the sum of the two flux integrals.
Using this, observation,
\[ \int_S \text{curl} (\mathbf{F}) \cdot \mathbf{n} \, dS + \int_B \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & -1 \end{pmatrix} : \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & -1 \end{pmatrix} \, dS = 0 \]
where \( B \) is the base consisting of the ellipse
\[ \frac{x^2}{4} + \frac{y^2}{9} \leq 1 \]
It has area \( \pi (2)(3) = 6\pi \) and so
\[ \int_S \text{curl} (\mathbf{F}) \cdot \mathbf{n} \, dS + 6\pi = 0 \]
which gives the same answer as what was just done.

**Example 28.3.6** Let \( \mathbf{F} (x, y, z) = (xy, x + y, zy) \). Let \( S \) be the surface which is formed by the intersection of the plane \( x + 2y + 2z = 1 \) with the cylinder \( x^2 + y^2 \leq 1 \). Let \( C \) be the intersection of the plane with the cylinder oriented counter clockwise when viewed from high on the \( z \) axis. Find \( \int_C \mathbf{F} \cdot d\mathbf{r} \).

You ought to use Stoke’s theorem. \( \nabla \times (xy, x + y, zy) = \left( z, \ 0, \ 1 - x \right) \). Then the line integral equals \( \int_S \text{curl} (\mathbf{F}) \cdot \mathbf{n} \, dS \). What is \( \mathbf{n} \)? It ought to point up and we can get it as follows.

\[ \mathbf{n} = \frac{(1, 2, 2)}{3} \]

What is \( dS \)? it is given by \( dS = \sqrt{1 + \left( \frac{1}{2} \right)^2 + 1} \, dxdy \) because \( z = \frac{1 - 2y - x}{2} \). Then letting \( B \) denote the inside of the unit circle, we need to find

\[ \int_B \left( z, \ 0, \ 1 - x \right) \cdot \begin{pmatrix} 1, & 2, & 2 \\ 3, & 3, & 3 \end{pmatrix} \sqrt{1 + \left( \frac{1}{2} \right)^2 + 1} \, dxdy \]

Of course you have to write the \( z \) in terms of \( x, y \). Thus the integral is

\[ \int_B \left( \frac{1 - 2y - x}{2}, \ 0, \ 1 - x \right) \cdot \begin{pmatrix} 1, & 2, & 2 \\ 3, & 3, & 3 \end{pmatrix} \frac{3}{2} dxdy \]

\[ = \int_B \left( \frac{5}{4} - \frac{1}{2} y - \frac{5}{4} x \right) dxdy \]

By symmetry considerations, this is just

\[ \int_B \frac{5}{4} dxdy = \frac{5}{4} \pi \]
28.3.3 The Meaning Of The Curl

Imagine a little circular disk of radius \( r \) denoted as \( D_r \) which is perpendicular to a unit vector \( n \). Letting the thumb of the right hand point in the direction \( n \) the fingers of the right hand “curl” in the direction of motion around this curve. Equivalently, if you walked around the bounding circle with your left hand over the surface, you would be traveling in the appropriate direction. Let \( \mathbf{F} \) be a vector field. Then by Stokes theorem, you have

\[
\int_{D_r} \nabla \times \mathbf{F} \cdot n \, dS = \int_{\partial D_r} \mathbf{F} \cdot d\mathbf{r}
\]

Then divide by the area of the little disk \( A_r = \pi r^2 \). You then get

\[
\frac{1}{A_r} \int_{D_r} \nabla \times \mathbf{F} \cdot n \, dS = \frac{1}{A_r} \int_{\partial D_r} \mathbf{F} \cdot d\mathbf{r}
\]

The line integral on the right is called the circulation of \( \mathbf{F} \) around the circle. Then you take a limit as \( r \to 0 \) to conclude that

\[
\nabla \times \mathbf{F} \cdot n = \lim_{r \to 0} \frac{1}{A_r} \int_{\partial D_r} \mathbf{F} \cdot d\mathbf{r}
\]

Thus the curl is a measure of the circulation of the vector field about the given vector \( n \) in the limit as this disk gets smaller and smaller. Sometimes people like to think of this as the tendency of the vector field to spin or rotate a little paddle wheel which is attached to the base of the vector \( n \) when it is inserted into the vector field.

28.3.4 Conservative Vector Fields

**Definition 28.3.7** A vector field, \( \mathbf{F} \) defined in a three dimensional region is said to be **conservative** \(^2\) if for every piecewise smooth closed curve, \( C \), it follows \( \int_C \mathbf{F} \cdot d\mathbf{R} = 0 \).

**Definition 28.3.8** Let \((x, p_1, \cdots, p_n, y)\) be an ordered list of points in \( \mathbb{R}^p \). Let \( p(x, p_1, \cdots, p_n, y) \)

denote the piecewise smooth curve consisting of a straight line segment from \( x \) to \( p_1 \) and then the straight line segment from \( p_1 \) to \( p_2 \), \( \cdots \) and finally the straight line segment from \( p_n \) to \( y \). 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, \( x, y \in U \), 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 28.3.9** 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 \( x \) to \( y \) in \( U \), then

\[
\int_C \mathbf{F} \cdot d\mathbf{R} = \phi(y) - \phi(x).
\]

Thus the line integral is path independent in this case. This function, \( \phi \) is called a **scalar potential** for \( \mathbf{F} \).\(^2\)

\[^2\]There is no such thing as a liberal vector field.
28.3. STOKE’S THEOREM FROM GREEN’S THEOREM

**Proof:** To save space and fussing over things which are unimportant, denote by \( p(x_0, x) \) a polygonal curve from \( x_0 \) to \( x \). Thus the orientation is such that it goes from \( x_0 \) to \( x \). The curve \( p(x, x_0) \) denotes the same set of points but in the opposite order. Suppose first \( F \) is conservative. Fix \( x_0 \in U \) and let

\[
\phi(x) \equiv \int_{p(x_0,x)} F \cdot dR.
\]

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

\[
\int_{p(x_0, x)} F \cdot dR + \int_{q(x, x_0)} F \cdot dR = \int_{p(x_0, x)} F \cdot dR - \int_{q(x, x_0)} F \cdot dR
\]

which shows

\[
\int_{p(x_0, x)} F \cdot dR = \int_{q(x, x_0)} F \cdot dR
\]

and that \( \phi \) is well defined. For small \( t \),

\[
\frac{\phi(x + te_i) - \phi(x)}{t} = \frac{\int_{p(x_0, x + te_i)} F \cdot dR - \int_{p(x_0, x)} F \cdot dR}{t}
\]

\[
= \frac{\int_{p(x_0, x)} F \cdot dR + \int_{p(x, x + te_i)} F \cdot dR - \int_{p(x_0, x)} F \cdot dR}{t}
\]

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 \( x + te_i \) is also contained in \( U \) and so one can take \( p(x, x + te_i)(s) = x + s(te_i) \) for \( s \in [0, 1] \). Therefore, the above difference quotient reduces to

\[
\frac{1}{t} \int_0^1 F(x + s(te_i)) \cdot te_i \, ds = \int_0^1 F_i(x + s(te_i)) \, ds
\]

\[
= F_i(x + s_i(te_i))
\]

by the mean value theorem for integrals. Here \( s_i \) is some number between 0 and 1. By continuity of \( F \), this converges to \( F_i(x) \) as \( t \to 0 \). Therefore, \( \nabla \phi = F \) as claimed.

Conversely, if \( \nabla \phi = F \), then if \( R : [a, b] \to \mathbb{R}^p \) is any \( C^1 \) curve joining \( x \) to \( y \),

\[
\int_a^b F(R(t)) \cdot R'(t) \, dt = \int_a^b \nabla \phi(R(t)) \cdot R'(t) \, dt
\]

\[
= \int_a^b \frac{d}{dt} \phi(R(t)) \, dt
\]

\[
= \phi(R(b)) - \phi(R(a))
\]

\[
= \phi(y) - \phi(x)
\]

and this verifies 28.3 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 28.3 implies the line integral over any closed curve equals zero because the starting and ending points of such a curve are the same. \( \blacksquare \)
**Example 28.3.10** Let $F(x, y, z) = (\cos x - yz \sin(xz), \cos(xz), -yx \sin(xz))$. Let $C$ be a piecewise smooth curve which goes from $(\pi, 1, 1)$ to $(\frac{\pi}{2}, 3, 2)$. Find $\int_C F \cdot dR$.

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 $\nabla \phi = F$. Such a function satisfies

$$\phi_x = \cos x - y (\sin xz) z$$

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 28.3.11** Let $F(x, y, z) = (x, y^2 x, z)$. Find a scalar potential for $F$ 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 28.3.12** 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, $C$. 
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 28.3.13** For a three dimensional simply connected open set, V and F a C¹ 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

\[
0 = \int_S \nabla \times F \cdot \mathbf{n} dA = \int_C F \cdot d\mathbf{R}.
\]

Thus \( F \) is conservative.

**Example 28.3.14** Determine whether the vector field,

\[
(4x^3 + 2(\cos(x^2 + z^2))) \mathbf{x}, 1, 2(\cos(x^2 + z^2)) \mathbf{z}
\]

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.

\[
\begin{vmatrix}
\mathbf{i} & \mathbf{j} & \mathbf{k} \\
\partial_x & \partial_y & \partial_z \\
4x^3 + 2(\cos(x^2 + z^2)) & 1 & 2(\cos(x^2 + z^2)) \mathbf{z}
\end{vmatrix}
\]

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 \( \phi_z = 2(\cos(x^2 + z^2)) \mathbf{z} \) 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 \( 4x^3 + 2(\cos(x^2 + z^2)) \mathbf{x} = g'(x) + 2x \cos(x^2 + z^2) \) and so it suffices to take \( g(x) = x^4 \). Hence \( \phi(x, y, z) = y + x^4 + \sin(x^2 + z^2) \).

### 28.3.5 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 \( Pdx + Qdy + Rdz \) is exact. Some people like to say things in terms of vector fields and some say it in terms of differential forms. In Example 28.3.13, the differential form \( (4x^3 + 2(\cos(x^2 + z^2)) \mathbf{x}) dx + dy + (2(\cos(x^2 + z^2)) \mathbf{z}) dz \) is exact.

### 28.4 When Do You Use What?

As a review, the following table indicates when you use various integral formulas.

| Surface in \( \mathbb{R}^2 \), closed curve for boundary | Green’s theorem | \( \int_U (Q_x - P_y) \, dA = \int_C P \, dx + Q \, dy \) |
| Surface in \( \mathbb{R}^3 \), closed curve for boundary | Stoke’s theorem | \( \int_S \nabla \times F \cdot \mathbf{n} \, dS = \int_C F \cdot d\mathbf{r} \) |
| Closed surface in \( \mathbb{R}^3 \) | Divergence theorem | \( \int_V \nabla \cdot F \, dV = \int_{\partial V} F \cdot \mathbf{n} \, dS \) |
CHAPTER 28. STOKES AND GREEN’S THEOREMS

<table>
<thead>
<tr>
<th>Theorem</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green’s theorem</td>
<td>Surface (Flux) integral = Circulation integral</td>
</tr>
<tr>
<td>Stoke’s Theorem</td>
<td>Surface (Flux) integral = Circulation integral</td>
</tr>
<tr>
<td>Divergence theorem</td>
<td>Volume integral = Flux integral</td>
</tr>
</tbody>
</table>

Note that there are really only two kinds of theorems here. Actually there is really only one but we will not get in to this. Recall that Green’s theorem is really a special case of Stoke’s theorem. In fact $\nabla \times (P(x, y), Q(x, y), 0) = (Q_y - P_y) \mathbf{k}$ and $\mathbf{k}$ is the appropriate unit normal to the surface in $\mathbb{R}^2$. Thus $(Q_x - P_y) = \nabla \times (P(x, y), Q(x, y), 0) \cdot \mathbf{k}$ with $\mathbf{k}$ normal to the surface. Thus the two top theorems, Green’s theorem and Stoke’s theorem are both relating a flux integral over a surface to a circulation integral along the bounding curve for the surface. The divergence theorem relates a volume integral and a flux integral over the bounding surface. Note how they all involve a flux integral over a surface. With the top two, they relate the flux integral to a line integral and in the case of the divergence theorem, the flux integral relates to a volume integral. In the top two, the surface is typically not closed. In the bottom, the surface is closed.

28.5 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

\[
\begin{align*}
\nabla \times \mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{B}}{\partial t} &= 0 \\
\nabla \cdot \mathbf{E} &= 4\pi \rho \\
\nabla \times \mathbf{B} - \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t} &= \frac{4\pi}{c} \mathbf{f} \\
\n\nabla \cdot \mathbf{B} &= 0
\end{align*}
\]  

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 (25.3) $\mathbf{B}$ has a vector potential, $\mathbf{A}_1$ such that $\mathbf{B} = \nabla \times \mathbf{A}_1$. Now go to (25.3) 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}.
\]

(28.9)

Now suppose $\phi$ is a time dependent scalar field satisfying

\[
\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = \frac{1}{c} \frac{\partial \psi_1}{\partial t} - \nabla \cdot \mathbf{A}_1.
\]

(28.10)

Next define

\[
\mathbf{A} \equiv \mathbf{A}_1 + \nabla \phi, \quad \psi \equiv \psi_1 + \frac{1}{c} \frac{\partial \phi}{\partial t}.
\]

(28.11)
28.5. Maxwell’s Equations and the Wave Equation

Therefore, in terms of the new variables, \((28.10)\) becomes

\[
\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = \frac{1}{c} \left( \frac{\partial \psi}{\partial t} - \frac{1}{c} \frac{\partial^2 \phi}{\partial t^2} \right) - \nabla \cdot A + \nabla^2 \phi
\]

which yields

\[
0 = \frac{\partial \psi}{\partial t} - c \nabla \cdot A. \quad (28.12)
\]

Then it follows from Theorem 23.3.7 on Page 487 that \(A\) is also a vector potential for \(B\). That is

\[
\nabla \times A = B. \quad (28.13)
\]

From \((28.11)\)

\[
\nabla \left( \psi - \frac{1}{c} \frac{\partial \phi}{\partial t} \right) = E + \frac{1}{c} \left( \frac{\partial A}{\partial t} - \nabla \frac{\partial \phi}{\partial t} \right)
\]

and so

\[
\nabla \psi = E + \frac{1}{c} \frac{\partial A}{\partial t}. \quad (28.14)
\]

Using \((28.7)\) and \((28.14)\),

\[
\nabla \times (\nabla \times A) - \frac{1}{c} \frac{\partial}{\partial t} \left( \nabla \psi - \frac{1}{c} \frac{\partial A}{\partial t} \right) = \frac{4\pi}{c} f. \quad (28.15)
\]

Now from Theorem 23.3.7 on Page 487 this implies

\[
\nabla (\nabla \cdot A) - \nabla^2 A - \nabla \left( \frac{1}{c} \frac{\partial \psi}{\partial t} \right) + \frac{1}{c^2} \frac{\partial^2 A}{\partial t^2} = \frac{4\pi}{c} f
\]

and using \((28.12)\), this gives

\[
\frac{1}{c^2} \frac{\partial^2 A}{\partial t^2} - \nabla^2 A = \frac{4\pi}{c} f. \quad (28.16)
\]

Also from \((28.7)\), \((28.11)\) and \((28.12)\),

\[
\nabla^2 \psi = \nabla \cdot E + \frac{1}{c} \frac{\partial}{\partial t} (\nabla \cdot A)
\]

\[
= 4\pi \rho + \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2}
\]

and so

\[
\frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} - \nabla^2 \psi = -4\pi \rho. \quad (28.17)
\]

This is very interesting. If a solution to the wave equations, \((28.17)\) and \((28.16)\) can be found along with a solution to \((28.12)\), then letting the magnetic field be given by \((28.13)\) and letting \(E\) be given by \((28.14)\) 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 \((28.12)\) to hold.
Part X

Some Iterative Techniques For Linear Algebra
Chapter 29

Iterative Methods For Linear Systems

Consider the problem of solving the equation

\[ Ax = b \]  \hspace{1cm} (29.1)

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. I am following the treatment given to this subject by Nobel and Daniel [20].

There are two main methods which are used to obtain solutions iteratively, the Jacobi method and the Gauss Seidel method. I will illustrate with an example and then describe the method precisely.

### 29.1 Jacobi Method

**Example 29.1.1** Use the Jacobi method to find the solutions to the following system of equations.

\[
\begin{align*}
7x + y &= 11 \\
x - 5y &= 7
\end{align*}
\]

It is profoundly stupid to use the Jacobi method on such a \( 2 \times 2 \) system. You should simply use row operations. If you do, the solution is \( \{ y = -\frac{19}{18}, x = \frac{31}{18} \} \). In terms of decimals this is \( \{ y = -1.05555556, x = 1.72222222 \} \). Now I will proceed to show how to use the Jacobi method to also find this solution.

Here are steps which describe the Jacobi method. You write the system as

\[
\begin{pmatrix} 7 & 1 \\ 1 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 11 \\ 7 \end{pmatrix}
\]

Next you split the matrix as follows

\[
\begin{pmatrix} 7 & 0 \\ 0 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 11 \\ 7 \end{pmatrix}
\]
That is you write the matrix as the sum of a diagonal matrix plus the off diagonal terms. Then if you have a solution, you would need

\[
\begin{pmatrix}
7 & 0 \\
0 & -5
\end{pmatrix}
\begin{pmatrix}
x \\
y
\end{pmatrix} =
-\begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}
\begin{pmatrix}
x \\
y
\end{pmatrix} + \begin{pmatrix}
11 \\
7
\end{pmatrix}
\]

You could write this as

\[
\begin{pmatrix}
x \\
y
\end{pmatrix} =
-\begin{pmatrix}
7 & 0 \\
0 & -5
\end{pmatrix}^{-1}
\begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}
\begin{pmatrix}
x \\
y
\end{pmatrix} + \begin{pmatrix}
7 & 0 \\
0 & -5
\end{pmatrix}^{-1}
\begin{pmatrix}
11 \\
7
\end{pmatrix}
\]

\[
= \left( \frac{1}{7} \right)
\begin{pmatrix}
x \\
y
\end{pmatrix} + \begin{pmatrix}
\frac{11}{7} \\
\frac{5}{7}
\end{pmatrix}
\]

This suggests a way to approach the problem through a process of iterations. You pick an initial guess for \((x, y)\) say \((0, 0)\). (It really doesn’t matter what you pick. When the method works it will do so for any initial choice.) Call this initial guess

\[
\begin{pmatrix}
x_0 \\
y_0
\end{pmatrix}
\]

and then you obtain the next guess, \(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}\) as follows

\[
\begin{pmatrix}
x_1 \\
y_1
\end{pmatrix} = \left( \frac{1}{7} \right)
\begin{pmatrix}
x_0 \\
y_0
\end{pmatrix} + \begin{pmatrix}
\frac{11}{7} \\
\frac{5}{7}
\end{pmatrix}
\]

Then to get the next guess you do the same thing.

\[
\begin{pmatrix}
x_2 \\
y_2
\end{pmatrix} = \left( \frac{1}{7} \right)
\begin{pmatrix}
x_1 \\
y_1
\end{pmatrix} + \begin{pmatrix}
\frac{11}{7} \\
\frac{5}{7}
\end{pmatrix}
\]

Continuing this way, this method hopefully will give guesses which are increasingly close to the true solution. Let’s apply this to this example.

\[
\begin{pmatrix}
x_1 \\
y_1
\end{pmatrix} = \left( \frac{1}{7} \right)
\begin{pmatrix}
0 \\
1
\end{pmatrix} \begin{pmatrix}
1.57142857 \\
-1.4
\end{pmatrix} + \begin{pmatrix}
\frac{11}{7} \\
\frac{5}{7}
\end{pmatrix}
\]

Now you find the next guess.

\[
\begin{pmatrix}
x_2 \\
y_2
\end{pmatrix} = \left( \frac{1}{7} \right)
\begin{pmatrix}
1.57142857 \\
-1.4
\end{pmatrix} + \begin{pmatrix}
\frac{11}{7} \\
\frac{5}{7}
\end{pmatrix}
\]

Things are still changing so I will try the next guess.

\[
\begin{pmatrix}
x_3 \\
y_3
\end{pmatrix} = \left( \frac{1}{7} \right)
\begin{pmatrix}
1.77142857 \\
-1.08571429
\end{pmatrix} + \begin{pmatrix}
\frac{11}{7} \\
\frac{5}{7}
\end{pmatrix}
\]

\[
= \begin{pmatrix}
1.72653061 \\
-1.04571429
\end{pmatrix}
\]
29.1. JACOBI METHOD

Let’s do another iteration.

\[
\begin{pmatrix}
x_4 \\
y_4
\end{pmatrix} = \begin{pmatrix}
0 & -\frac{1}{7} \\
\frac{1}{5} & 0
\end{pmatrix} \begin{pmatrix}
1.72653061 \\
-1.04571429
\end{pmatrix} + \begin{pmatrix}
\frac{11}{7} \\
-\frac{7}{5}
\end{pmatrix}
\]

\[
= \begin{pmatrix}
1.72081633 \\
-1.05469388
\end{pmatrix}.
\]

This should be pretty close because the guesses are not changing much from one to the next. The exact solution was

\[
\{y = -1.05555556, x = 1.72222222\}
\]

Actually, you don’t do it this way. The following gives the way a computer would do it. You do not invert the matrix as I did. However, for the purposes of illustration and for small systems there is no harm in doing it as I did above, especially since for small systems of equations it is a stupid idea to use an iterative method in the first place.

Definition 29.1.2 The Jacobi iterative technique, also called the method of simultaneous corrections is defined as follows. Let \(x^1\) be an initial vector, say the zero vector or some other vector. The method generates a succession of vectors, \(x^2, x^3, x^4, \cdots\) and hopefully this sequence of vectors will converge to the solution to 29.1. The vectors in this list are called iterates and they are obtained according to the following procedure. Letting \(A = (a_{ij})\),

\[
a_{ij}x_i^{r+1} = -\sum_{j \neq i} a_{ij}x_j^r + b_i. \tag{29.2}
\]

In terms of matrices, letting

\[
A = \begin{pmatrix}
* & \cdots & * \\
\vdots & \ddots & \vdots \\
* & \cdots & *
\end{pmatrix}
\]

The iterates are defined as

\[
\begin{pmatrix}
x_1^{r+1} \\
x_2^{r+1} \\
\vdots \\
x_n^{r+1}
\end{pmatrix} = -A \begin{pmatrix}
x_1^r \\
x_2^r \\
\vdots \\
x_n^r
\end{pmatrix} + \begin{pmatrix}
b_1 \\
b_2 \\
\vdots \\
b_n
\end{pmatrix} \tag{29.3}
\]

The matrix on the left in 29.3 is obtained by retaining the main diagonal of \(A\) and setting every other entry equal to zero. The matrix on the right in 29.3 is obtained from \(A\) by setting every diagonal entry equal to zero and retaining all the other entries unchanged.

Example 29.1.3 Use the Jacobi method to solve the system

\[
\begin{pmatrix}
3 & 1 & 0 & 0 \\
1 & 4 & 1 & 0 \\
0 & 2 & 5 & 1 \\
0 & 0 & 2 & 4
\end{pmatrix} \begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{pmatrix} = \begin{pmatrix}
1 \\
2 \\
3 \\
4
\end{pmatrix}
\]
Of course this is solved most easily using row reductions. The Jacobi method is useful when the matrix is 1000×1000 or larger. This example is just to illustrate how the method works. First let’s solve it using row operations. The augmented matrix is
\[
\begin{pmatrix}
3 & 1 & 0 & 0 & 1 \\
1 & 4 & 1 & 0 & 2 \\
0 & 2 & 5 & 1 & 3 \\
0 & 0 & 2 & 4 & 4
\end{pmatrix}
\]

The row reduced echelon form is
\[
\begin{pmatrix}
1 & 0 & 0 & 0 & 6/29 \\
0 & 1 & 0 & 0 & 11/29 \\
0 & 0 & 1 & 0 & 8/29 \\
0 & 0 & 0 & 1 & 25/29
\end{pmatrix}
\]

which in terms of decimals is approximately equal to
\[
\begin{pmatrix}
1.0 & 0 & 0 & 0 & .206 \\
0 & 1.0 & 0 & 0 & .379 \\
0 & 0 & 1.0 & 0 & .275 \\
0 & 0 & 0 & 1.0 & .862
\end{pmatrix}
\]

In terms of the matrices, the Jacobi iteration is of the form
\[
\begin{pmatrix}
x_1^{r+1} \\
x_2^{r+1} \\
x_3^{r+1} \\
x_4^{r+1}
\end{pmatrix} = - \begin{pmatrix}
0 & 1/3 & 0 & 0 \\
1/4 & 0 & 1/4 & 0 \\
2/5 & 0 & 1/5 & 0 \\
0 & 1/2 & 0 & 1
\end{pmatrix} \begin{pmatrix}
x_1^r \\
x_2^r \\
x_3^r \\
x_4^r
\end{pmatrix} + \begin{pmatrix}
1 \\
2 \\
3 \\
4
\end{pmatrix}.
\]

Multiplying by the inverse of the matrix on the left, this iteration reduces to
\[
\begin{pmatrix}
x_1^{r+1} \\
x_2^{r+1} \\
x_3^{r+1} \\
x_4^{r+1}
\end{pmatrix} = - \begin{pmatrix}
0 & 1/3 & 0 & 0 \\
1/4 & 0 & 1/4 & 0 \\
2/5 & 0 & 1/5 & 0 \\
0 & 1/2 & 0 & 1
\end{pmatrix} \begin{pmatrix}
x_1^r \\
x_2^r \\
x_3^r \\
x_4^r
\end{pmatrix} + \begin{pmatrix}
1/3 \\
1/3 \\
1/3 \\
1/3
\end{pmatrix}.
\]

Now iterate this starting with
\[
x^1 = \begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}.
\]

Thus
\[
x^2 = - \begin{pmatrix}
0 & 1/3 & 0 & 0 \\
1/4 & 0 & 1/4 & 0 \\
2/5 & 0 & 1/5 & 0 \\
0 & 1/2 & 0 & 1
\end{pmatrix} \begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix} + \begin{pmatrix}
1/3 \\
1/3 \\
1/3
\end{pmatrix} = \begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}.
\]

\[\text{You certainly would not compute the inverse in solving a large system. This is just to show you how the method works for this simple example. You would use the first description in terms of indices.}\]
Then

\[
x^3 = - \begin{pmatrix} 0 & 1/3 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 2/3 \\ 4/3 \\ 0 \\ 2/5 \\ 0 \\ 1/2 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} .166 \\ .26 \\ .7 \end{pmatrix}
\]

\[
x^4 = - \begin{pmatrix} 0 & 1/3 & 0 & 0 \end{pmatrix} \begin{pmatrix} .166 \\ .26 \\ .2 \\ .7 \end{pmatrix} + \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} .24 \\ .4085 \\ .356 \\ .9 \end{pmatrix}
\]

You can keep going like this. Recall the solution is approximately equal to

\[
\begin{pmatrix} .206 \\ .379 \\ .275 \\ .862 \end{pmatrix}
\]

so you see that with no care at all and only 6 iterations, an approximate solution has been obtained which is not too far off from the actual solution.

It is important to realize that a computer would use \texttt{GAUSS} directly. Indeed, writing the problem in terms of matrices as I have done above destroys every benefit of the method. However, it makes it a little easier to see what is happening and so this is why I have presented it in this way.

### 29.2 Gauss Seidel Method

**Example 29.2.1** Solve the following system of equations using the Gauss Seidel method. It is the same example as in Example 29.1.1.

\[
\begin{aligned}
7x + y &= 11 \\
x - 5y &= 7
\end{aligned}
\]
The solution to this system is: \( \{ y = -1.055\ 555\ 56, x = 1.722\ 222\ 22 \} \). Now I will use the Gauss Seidel method to get this solution. The system is of the form

\[
\begin{pmatrix} 7 & 0 \\ 1 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 11 \\ 7 \end{pmatrix}
\]

Note the difference! Here you split the matrix differently. Then the iteration scheme is just as before,

\[
\begin{pmatrix} 7 & 0 \\ 1 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = -\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 11 \\ 7 \end{pmatrix}
\]

and so the solution satisfies

\[
\begin{pmatrix} x \\ y \end{pmatrix} = -\begin{pmatrix} 7 & 0 \\ 1 & -5 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 7 & 0 \\ 1 & -5 \end{pmatrix}^{-1} \begin{pmatrix} 11 \\ 7 \end{pmatrix}
\]

The corresponding iteration scheme yields

\[
\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1}{7} \\ 0 & -\frac{1}{35} \end{pmatrix} \begin{pmatrix} x_n \\ y_n \end{pmatrix} + \begin{pmatrix} \frac{11}{7} \\ -\frac{38}{35} \end{pmatrix}
\]

Starting with an initial guess of \( x_0 = y_0 = 0 \), consider the following iterations.

\[
\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1}{7} \\ 0 & -\frac{1}{35} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{11}{7} \\ -\frac{38}{35} \end{pmatrix} = \begin{pmatrix} 1.571\ 428\ 57 \\ -1.085\ 714\ 29 \end{pmatrix}
\]

\[
\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1}{7} \\ 0 & -\frac{1}{35} \end{pmatrix} \begin{pmatrix} 1.571\ 428\ 57 \\ -1.085\ 714\ 29 \end{pmatrix} + \begin{pmatrix} \frac{11}{7} \\ -\frac{38}{35} \end{pmatrix} = \begin{pmatrix} 1.726\ 530\ 61 \\ -1.054\ 693\ 88 \end{pmatrix}
\]

\[
\begin{pmatrix} x_3 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1}{7} \\ 0 & -\frac{1}{35} \end{pmatrix} \begin{pmatrix} 1.726\ 530\ 61 \\ -1.054\ 693\ 88 \end{pmatrix} + \begin{pmatrix} \frac{11}{7} \\ -\frac{38}{35} \end{pmatrix} = \begin{pmatrix} 1.722\ 099\ 13 \\ -1.055\ 580\ 17 \end{pmatrix}
\]

These guesses are pretty close so it seems this should be close. Note the exact solution is \( \{ y = -1.055\ 555\ 56, x = 1.722\ 222\ 22 \} \). I think you can see this method worked a little better than the Jacobi method although both are pretty good.

The following is the precise description of the method. As before, you don’t write out the matrices and invert that matrix like above.
Definition 29.2.2 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

\[
\sum_{j=1}^{i} a_{ij} x_j^{r+1} = - \sum_{j=i+1}^{n} a_{ij} x_j^r + b_i. \tag{29.5}
\]

In terms of matrices, letting

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

The iterates are defined as

\[
\begin{pmatrix}
\ast & 0 & \cdots & 0 \\
\ast & \ast & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ast \\
\ast & \cdots & \ast & \ast
\end{pmatrix}
\begin{pmatrix}
x_1^{r+1} \\
x_2^{r+1} \\
\vdots \\
x_n^{r+1}
\end{pmatrix}
= -\begin{pmatrix}
0 & \ast & \cdots & \ast \\
0 & 0 & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ast \\
0 & \cdots & 0 & 0
\end{pmatrix}
\begin{pmatrix}
x_1^r \\
x_2^r \\
\vdots \\
x_n^r
\end{pmatrix}
+ \begin{pmatrix}
b_1 \\
b_2 \\
\vdots \\
b_n
\end{pmatrix}. \tag{29.6}
\]

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 (29.5) 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 29.2.3 Use the Gauss Seidel method to solve the system

\[
\begin{pmatrix}
3 & 1 & 0 & 0 \\
1 & 4 & 1 & 0 \\
0 & 2 & 5 & 1 \\
0 & 0 & 2 & 4
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{pmatrix}
= \begin{pmatrix}
1 \\
2 \\
3 \\
4
\end{pmatrix}
\]

In terms of matrices, this procedure is

\[
\begin{pmatrix}
3 & 0 & 0 & 0 \\
1 & 4 & 0 & 0 \\
0 & 2 & 5 & 0 \\
0 & 0 & 2 & 4
\end{pmatrix}
\begin{pmatrix}
x_1^{r+1} \\
x_2^{r+1} \\
x_3^{r+1} \\
x_4^{r+1}
\end{pmatrix}
= -\begin{pmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
x_1^r \\
x_2^r \\
x_3^r \\
x_4^r
\end{pmatrix}
+ \begin{pmatrix}
1 \\
2 \\
3 \\
4
\end{pmatrix}
\].
Multiplying by the inverse of the matrix on the left\(^2\) this yields

\[
\begin{pmatrix}
  x^r_1 + 1 \\
  x^r_2 + 1 \\
  x^r_3 + 1 \\
  x^r_4 + 1
\end{pmatrix} = -\begin{pmatrix}
  0 & \frac{1}{3} & 0 & 0 \\
  0 & -\frac{1}{12} & \frac{1}{4} & 0 \\
  0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\
  0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10}
\end{pmatrix} \begin{pmatrix}
  x^r_1 \\
  x^r_2 \\
  x^r_3 \\
  x^r_4
\end{pmatrix} + \begin{pmatrix}
  \frac{1}{3} \\
  \frac{5}{12} \\
  \frac{13}{30} \\
  \frac{47}{60}
\end{pmatrix}
\]

As before, I will be totally unoriginal in the choice of \(x^1\). Let it equal the zero vector. Therefore,

\[
x^2 = \begin{pmatrix}
  \frac{1}{3} \\
  \frac{5}{12} \\
  \frac{13}{30} \\
  \frac{47}{60}
\end{pmatrix}
\]

Now

\[
x^3 = -\begin{pmatrix}
  0 & \frac{1}{3} & 0 & 0 \\
  0 & -\frac{1}{12} & \frac{1}{4} & 0 \\
  0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\
  0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10}
\end{pmatrix} \begin{pmatrix}
  x^2_1 \\
  x^2_2 \\
  x^2_3 \\
  x^2_4
\end{pmatrix} + \begin{pmatrix}
  \frac{1}{3} \\
  \frac{5}{12} \\
  \frac{13}{30} \\
  \frac{47}{60}
\end{pmatrix} = \begin{pmatrix}
  .194 \\
  .343 \\
  .306 \\
  .846
\end{pmatrix}
\]

It follows

\[
x^4 = -\begin{pmatrix}
  0 & \frac{1}{3} & 0 & 0 \\
  0 & -\frac{1}{12} & \frac{1}{4} & 0 \\
  0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\
  0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10}
\end{pmatrix} \begin{pmatrix}
  x^3_1 \\
  x^3_2 \\
  x^3_3 \\
  x^3_4
\end{pmatrix} + \begin{pmatrix}
  \frac{1}{3} \\
  \frac{5}{12} \\
  \frac{13}{30} \\
  \frac{47}{60}
\end{pmatrix} = \begin{pmatrix}
  .194 \\
  .343 \\
  .306 \\
  .846
\end{pmatrix}
\]

and so

\[
x^5 = -\begin{pmatrix}
  0 & \frac{1}{3} & 0 & 0 \\
  0 & -\frac{1}{12} & \frac{1}{4} & 0 \\
  0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\
  0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10}
\end{pmatrix} \begin{pmatrix}
  x^4_1 \\
  x^4_2 \\
  x^4_3 \\
  x^4_4
\end{pmatrix} + \begin{pmatrix}
  \frac{1}{3} \\
  \frac{5}{12} \\
  \frac{13}{30} \\
  \frac{47}{60}
\end{pmatrix} = \begin{pmatrix}
  .219 \\
  .36875 \\
  .2833 \\
  .85835
\end{pmatrix}
\]

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.

\(^2\)As in the case of the Jacobi iteration, the computer would not do this. It would use the iteration procedure in terms of the entries of the matrix directly. Otherwise all benefit to using this method is lost.
Example 29.2.4 Use the Gauss Seidel method to solve the system

\[
\begin{bmatrix}
1 & 4 & 0 & 0 \\
1 & 4 & 1 & 0 \\
0 & 2 & 5 & 1 \\
0 & 0 & 2 & 4
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{bmatrix}
= 
\begin{bmatrix}
1 \\
2 \\
3 \\
4
\end{bmatrix}
\]

The exact solution is given by doing row operations on the augmented matrix. When this is done the row echelon form is

\[
\begin{bmatrix}
1 & 0 & 0 & 0 & 6 \\
0 & 1 & 0 & 0 & \frac{5}{4} \\
0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & \frac{1}{2}
\end{bmatrix}
\]

and so the solution is approximately

\[
\begin{bmatrix}
6 \\
-\frac{5}{4} \\
1 \\
\frac{1}{2}
\end{bmatrix}
= 
\begin{bmatrix}
6.0 \\
-1.25 \\
1.0 \\
0.5
\end{bmatrix}
\]

The Gauss Seidel iterations are of the form

\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
1 & 4 & 0 & 0 \\
0 & 2 & 5 & 0 \\
0 & 0 & 2 & 4
\end{bmatrix}
\begin{bmatrix}
x_1^{r+1} \\
x_2^{r+1} \\
x_3^{r+1} \\
x_4^{r+1}
\end{bmatrix}
= 
\begin{bmatrix}
0 & 4 & 0 & 0 \\
0 & -1 & 1/4 & 0 \\
0 & 2/5 & -1/10 & 1/5 \\
0 & -1/5 & -1/20 & -1/10
\end{bmatrix}
\begin{bmatrix}
x_1^r \\
x_2^r \\
x_3^r \\
x_4^r
\end{bmatrix}
+ 
\begin{bmatrix}
1 \\
2 \\
3 \\
4
\end{bmatrix}
\]

and so, multiplying by the inverse of the matrix on the left, the iteration reduces to the following in terms of matrix multiplication.

\[
\begin{bmatrix}
x_1^{r+1} \\
x_2^{r+1} \\
x_3^{r+1} \\
x_4^{r+1}
\end{bmatrix}
= 
\begin{bmatrix}
0 & 4 & 0 & 0 \\
0 & -1 & 1/4 & 0 \\
0 & 2/5 & -1/10 & 1/5 \\
0 & -1/5 & -1/20 & -1/10
\end{bmatrix}
\begin{bmatrix}
x_1^r \\
x_2^r \\
x_3^r \\
x_4^r
\end{bmatrix}
+ 
\begin{bmatrix}
1 \\
2 \\
3 \\
4
\end{bmatrix}
\]

This time, I will pick an initial vector close to the answer. Let

\[
x_1 = \begin{bmatrix}
6 \\
-1 \\
1 \\
\frac{1}{2}
\end{bmatrix}
\]

This is very close to the answer. Now lets see what the Gauss Seidel iteration does to it.

\[
x_2 = 
\begin{bmatrix}
0 & 4 & 0 & 0 \\
0 & -1 & 1/4 & 0 \\
0 & 2/5 & -1/10 & 1/5 \\
0 & -1/5 & -1/20 & -1/10
\end{bmatrix}
\begin{bmatrix}
6 \\
-1 \\
1 \\
\frac{1}{2}
\end{bmatrix}
+ 
\begin{bmatrix}
1 \\
2 \\
3 \\
4
\end{bmatrix}
= 
\begin{bmatrix}
5.0 \\
-1.0 \\
1.0 \\
0.55
\end{bmatrix}
\]
You can’t expect to be real close after only one iteration. Let’s do another.

\[
x^3 = -
\begin{pmatrix}
  0 & 4 & 0 & 0 \\
  0 & -1 & 1 & 0 \\
  2 & 5 & -1 & \frac{1}{10} \\
  0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10}
\end{pmatrix}
\begin{pmatrix}
  5.0 \\
  -1.0 \\
  .9 \\
  .55
\end{pmatrix}
+ 
\begin{pmatrix}
  1 \\
  \frac{1}{4} \\
  \frac{1}{2} \\
  \frac{3}{4}
\end{pmatrix} = 
\begin{pmatrix}
  5.0 \\
  -.975 \\
  .88 \\
  .56
\end{pmatrix}
\]

\[
x^4 = -
\begin{pmatrix}
  0 & 4 & 0 & 0 \\
  0 & -1 & 1 & 0 \\
  2 & 5 & -1 & \frac{1}{10} \\
  0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10}
\end{pmatrix}
\begin{pmatrix}
  5.0 \\
  -.975 \\
  .88 \\
  .56
\end{pmatrix}
+ 
\begin{pmatrix}
  1 \\
  \frac{1}{4} \\
  \frac{1}{2} \\
  \frac{3}{4}
\end{pmatrix} = 
\begin{pmatrix}
  4.9 \\
  -.945 \\
  .866 \\
  .567
\end{pmatrix}
\]

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?

Both iterative procedures for solving

\[Ax = b\]  \hspace{1cm} (29.7)

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

\[x^{r+1} = -B^{-1}Cx^r + B^{-1}b\]  \hspace{1cm} (29.8)

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. Let \(\{\lambda_1, \ldots, \lambda_n\}\) be the eigenvalues.

**Definition 29.2.5** The spectral radius of a matrix, \(M\), denoted as \(\rho(M)\) is

\[\max \{|\lambda_1|, \ldots, |\lambda_n|\}.\]

That is it is the maximum of the absolute values of the eigenvalues of \(M\).

The following gives the condition under which any of these iterates as in \(29.8\) converge.

**Theorem 29.2.6** Suppose \(\rho(B^{-1}C) < 1\). Then the iterates in \(29.8\) converge to the unique solution of \(29.7\).

The following definition is useful.
Definition 29.2.7 Suppose $A$ is an $n \times n$ matrix. Then $A$ is said to be strictly diagonally dominant if for every $i$,

$$|A_{ii}| > \sum_{j \neq i} |A_{ij}|.$$  

That is, the absolute value of the entry in the $ii$th position is larger than the sum of the absolute values of all the other entries on the $i$th row.

Theorem 29.2.8 In either the Jacobi or the Gauss Seidel methods, if the matrix of coefficients which gets split to yield an iteration technique is strictly diagonally dominant, then the method converges. This means the iterates get close to the solution to the original system of equations as the iteration progresses.
Chapter 30

Iterative Methods For Finding Eigenvalues

Quiz

1. Let $F = (x, y, zx)$ and let $S$ be the surface having parametrization $r(u, v) = (uv, u + v, v)$ for $(u, v) \in [0, 1] \times [0, 2]$. Find the flux integral,

$$\int_S F \cdot n dS$$

where $n$ is the unit normal to the surface which has the same direction as the parametric normal, $N(u, v) = r_u \times r_v$.

2. Find the flux integral,

$$\int_S F \cdot n dS$$

of $F = \nabla \times G$ where $G(x, y, z) = \left(\sin (x^2 yz), \ln (x^4 + 7z^2 + 1), e^{x^2 + y^2} z^5 \sin (z)\right)$ on the level surface, $S$ given by the ellipsoid $x^2/6 + y^2/7 + z^2/2 = 1$.

3. Let $C$ be the oriented curve consisting of directed line segments which go from $(0, 0, 0)$ to $(3, 2, 1)$ to $(1, 2, 3)$ to $(33, 45.7, 67.23)$ and then to $(1, 1, 1)$. Find the line integral,

$$\int_C (2xy + 1) dx + (x^2 + 1) dy + 2 z dz.$$

4. Find the Laplacian of $x^3 - 3xy^2$.

5. Find the circulation density (curl) of the vector field $(x^2 y, yz, z + x)$.

30.1 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

\[
u_m = A^n u_1 = \sum_{k=1}^{n-1} c_k \lambda_k^m x_k + \lambda_n^m c_n x_n.
\]

(30.1)

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. We call this a scaling factor. 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 \( u_3 = Au_2 / S_3 \). Continue this way. The scaling just described does not destroy the relative insignificance of the term involving a sum in (30.1). 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 \( m \),

\[
u_m = \frac{\lambda_n^m c_n x_n}{S_2 S_3 \cdots S_m} + \text{(relatively insignificant term)}.
\]

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 \( \lambda_n \approx S_{m+1} \). 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 30.1.1** 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 \( u_1 = (1, \ldots, 1)^T \), 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.

\[
\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ 1 \end{pmatrix} = \begin{pmatrix} 22 \\ -8 \\ -6 \end{pmatrix}
\]

Then

\[
u_3 = \frac{1}{22} \begin{pmatrix} 22 \\ -8 \\ -6 \end{pmatrix} = \begin{pmatrix} \frac{1}{11} \\ -\frac{4}{11} \\ -\frac{3}{11} \end{pmatrix} = \begin{pmatrix} 1.0 \\ -0.363636 \\ -0.272727 \end{pmatrix}.
\]

Continue doing this

\[
\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -0.363636 \\ -0.272727 \end{pmatrix} = \begin{pmatrix} 7.0909091 \\ -4.3636364 \\ 1.6363637 \end{pmatrix}
\]

Then

\[
u_4 = \begin{pmatrix} 1.0 \\ -0.61538 \\ 0.23077 \end{pmatrix}
\]

So far the scaling factors are changing fairly noticeably so continue.

\[
\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -0.61538 \\ 0.23077 \end{pmatrix} = \begin{pmatrix} 16.154 \\ -7.3846 \\ -1.3846 \end{pmatrix}
\]

\[
u_5 = \begin{pmatrix} 1.0 \\ -0.45714 \\ -8.5713 \times 10^{-2} \end{pmatrix}
\]
\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
1.0 \\
-0.45714 \\
-8.5713 \times 10^{-2}
\end{pmatrix}
= 
\begin{pmatrix}
10.457 \\
-5.4857 \\
0.5143
\end{pmatrix}
\]

\[u_6 = \begin{pmatrix} 1.0 \\ -0.5246 \\ 4.9182 \times 10^{-2} \end{pmatrix}\]

\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
1.0 \\
-0.5246 \\
4.9182 \times 10^{-2}
\end{pmatrix}
= 
\begin{pmatrix}
12.885 \\
-6.2951 \\
-0.29515
\end{pmatrix}
\]

\[u_7 = \begin{pmatrix} 1.0 \\ -0.48856 \\ -2.2906 \times 10^{-2} \end{pmatrix}\]

\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
1.0 \\
-0.48856 \\
-2.2906 \times 10^{-2}
\end{pmatrix}
= 
\begin{pmatrix}
11.588 \\
-5.8626 \\
0.13736
\end{pmatrix}
\]

\[u_8 = \begin{pmatrix} 1.0 \\ -0.50592 \\ 1.1854 \times 10^{-2} \end{pmatrix}\]

\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
1.0 \\
-0.50592 \\ 1.1854 \times 10^{-2}
\end{pmatrix}
= 
\begin{pmatrix}
12.213 \\
-6.0711 \\
-7.1082 \times 10^{-2}
\end{pmatrix}
\]

\[u_9 = \begin{pmatrix} 1.0 \\ -0.4971 \\ -5.8202 \times 10^{-3} \end{pmatrix}\]

\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
1.0 \\
-0.4971 \\ -5.8202 \times 10^{-3}
\end{pmatrix}
= 
\begin{pmatrix}
11.895 \\
-5.9651 \\
3.4861 \times 10^{-2}
\end{pmatrix}
\]

\[u_{10} = \begin{pmatrix} 1.0 \\ -0.50148 \\ 2.9307 \times 10^{-3} \end{pmatrix}\]

\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
1.0 \\
-0.50148 \\ 2.9307 \times 10^{-3}
\end{pmatrix}
= 
\begin{pmatrix}
12.053 \\
-6.0176 \\
-1.7672 \times 10^{-2}
\end{pmatrix}
\]

\[u_{11} = \begin{pmatrix} 1.0 \\ -0.49926 \\ -1.4662 \times 10^{-3} \end{pmatrix}\]
30.1. THE POWER METHOD FOR EIGENVALUES

At this point, you could stop because the scaling factors are not changing by much. They went from 11.895 to 12.053. It looks like the eigenvalue is something like 12 which is in fact the case. The eigenvector is approximately \( u_{11} \). 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
\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
1 \\
-0.49926 \\
-1.4662 \times 10^{-3}
\end{pmatrix}
= \begin{pmatrix}
11.974 \\
-5.9912 \\
8.8386 \times 10^{-3}
\end{pmatrix}
\] (30.2)
and
\[
12.053 \begin{pmatrix}
1 \\
-0.49926 \\
-1.4662 \times 10^{-3}
\end{pmatrix}
= \begin{pmatrix}
12.053 \\
-6.0176 \\
-1.7672 \times 10^{-2}
\end{pmatrix}
\] (30.3)

30.1.1 Rayleigh Quotient

In the above procedure, you can sometimes estimate the eigenvalue a little differently. If \( Ax = \lambda x \) then
\[
\lambda = \frac{Ax \cdot x}{|x|^2}
\]
and so, in the method above, you might get an estimate for the eigenvalue in this way. The above is called the Rayleigh quotient. In [30.2] where an approximate eigenvector has been found, you could estimate the eigenvalue as
\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix}
\begin{pmatrix}
1 \\
-0.49926 \\
-1.4662 \times 10^{-3}
\end{pmatrix}
\cdot
\begin{pmatrix}
1 \\
-0.49926 \\
-1.4662 \times 10^{-3}
\end{pmatrix}
\]
\[= 1 + (-0.49926)^2 + (-1.4662 \times 10^{-3})^2
\]
= 11.978788

The scaling factor was 12.053 and the Rayleigh quotient gave 11.978788. I guess that at least in this case the scaling factor wins. Let’s look at a symmetric matrix. The book says the convergence of the Rayleigh quotients is about twice as fast as the scaling factors for symmetric matrices.

Example 30.1.2 Use the Rayleigh quotient with the power method to estimate the dominant eigenvalue for the matrix,
\[
\begin{pmatrix}
2 & 0 & 1 \\
0 & 3 & 0 \\
1 & 0 & 5
\end{pmatrix}
\]

It turns out that the eigenvalues of this matrix are \( 3, \frac{7}{2} + \frac{1}{2}\sqrt{13}, \frac{7}{2} - \frac{1}{2}\sqrt{13} \). In terms of decimals, 3, 5.30277564, 1.69722436, and so the dominant eigenvalue is 5.30277564.
Use the power method with an initial approximation \((1, 1, 1)^T\). Thus
\[
\begin{pmatrix}
2 & 0 & 1 \\
0 & 3 & 0 \\
1 & 0 & 5 \\
\end{pmatrix}
\begin{pmatrix}
1 \\
1 \\
1 \\
\end{pmatrix}
= \begin{pmatrix}
3.0 \\
3.0 \\
6.0 \\
\end{pmatrix}
\]
and so
\[
u_1 = \frac{1}{6} \begin{pmatrix}
3.0 \\
3.0 \\
6.0 \\
\end{pmatrix} = \begin{pmatrix}
.5 \\
.5 \\
1.0 \\
\end{pmatrix}
\]
Next iteration,
\[
\begin{pmatrix}
2 & 0 & 1 \\
0 & 3 & 0 \\
1 & 0 & 5 \\
\end{pmatrix}
\begin{pmatrix}
.5 \\
.5 \\
1.0 \\
\end{pmatrix}
= \begin{pmatrix}
2.0 \\
1.5 \\
5.5 \\
\end{pmatrix}
\]
Then
\[
u_2 = \frac{1}{5.5} \begin{pmatrix}
2.0 \\
1.5 \\
5.5 \\
\end{pmatrix} = \begin{pmatrix}
.363636364 \\
.272727273 \\
1.0 \\
\end{pmatrix}
\]
The scaling factor, 5.5 is an approximation to the dominant eigenvalue, 5.30277564. Let's see what is obtained from the Rayleigh quotient.
\[
\frac{ \begin{pmatrix}
2 & 0 & 1 \\
0 & 3 & 0 \\
1 & 0 & 5 \\
\end{pmatrix}
\begin{pmatrix}
.363636364 \\
.272727273 \\
1.0 \\
\end{pmatrix} \cdot \begin{pmatrix}
.363636364 \\
.272727273 \\
1.0 \\
\end{pmatrix}}{(.363636364)^2 + (.272727273)^2 + 1}
= 5.15068493
\]
This is slightly better than the scaling factor, 5.5. Let's do another iteration. Let's see if we get a dramatic increase in accuracy.
\[
\begin{pmatrix}
2 & 0 & 1 \\
0 & 3 & 0 \\
1 & 0 & 5 \\
\end{pmatrix}
\begin{pmatrix}
.363636364 \\
.272727273 \\
1.0 \\
\end{pmatrix}
= \begin{pmatrix}
1.72727273 \\
.818181819 \\
5.36363636 \\
\end{pmatrix}
\]
Now
\[
u_3 = \frac{1}{5.36363636} \begin{pmatrix}
1.72727273 \\
.818181819 \\
5.36363636 \\
\end{pmatrix} = \begin{pmatrix}
.322033899 \\
.152542373 \\
1.0 \\
\end{pmatrix}
\]
The scaling factor, 5.36363636 is an approximation to the dominant eigenvalue, 5.30277564. Let's try the Rayleigh quotient again. This gives
\[
\frac{ \begin{pmatrix}
2 & 0 & 1 \\
0 & 3 & 0 \\
1 & 0 & 5 \\
\end{pmatrix}
\begin{pmatrix}
.322033899 \\
.152542373 \\
1.0 \\
\end{pmatrix} \cdot \begin{pmatrix}
.322033899 \\
.152542373 \\
1.0 \\
\end{pmatrix}}{(.322033899)^2 + (.152542373)^2 + 1}
= 5.25414224
\]
The Rayleigh quotient is still just a little bit closer.
30.2 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 30.2.1 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! 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 $Ax_k = \lambda_k x_k$, 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

$$u_{k+1} = \frac{(A - \alpha I)^{-1}u_k}{S_k}.$$
where $S_k$ is the element of $(A - \alpha I)^{-1} u_k$ which has largest absolute value. From Lemma 30.2.1,

$$u_{k+1} = \frac{\sum_{j=1}^{m} a_j \left( \frac{1}{\lambda_j - \alpha} \right)^k x_j + \left( \frac{1}{\lambda_n - \alpha} \right)^k y}{S_2 \cdots S_k}$$

$$= \left( \frac{1}{\lambda_n - \alpha} \right)^k \frac{\sum_{j=1}^{m} a_j \left( \frac{\lambda_j - \alpha}{\lambda_n - \alpha} \right)^k x_j + y}{S_2 \cdots S_k}.$$

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

$$u_{k+1} = \left( \frac{1}{\lambda_n - \alpha} \right)^k \frac{E_k + y}{S_2 \cdots S_k} \equiv C_k (E_k + y)$$

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 $k$,

$$u_{k+1} - C_k y = C_k (E_k + y) \approx 0$$

and multiplying by $(A - \alpha I)^{-1}$ yields

$$(A - \alpha I)^{-1} u_{k+1} - (A - \alpha I)^{-1} C_k y = (A - \alpha I)^{-1} u_{k+1} - C_k \left( \frac{1}{\lambda_n - \alpha} \right)^k y$$

$$\approx (A - \alpha I)^{-1} u_{k+1} - \left( \frac{1}{\lambda_n - \alpha} \right) u_{k+1} \approx 0.$$ 

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

$$S_k u_{k+1} = (A - \alpha I)^{-1} u_k \approx (A - \alpha I)^{-1} (C_{k-1} y) = \frac{1}{\lambda_n - \alpha} (C_{k-1} y) \approx \frac{1}{\lambda_n - \alpha} u_k.$$

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

$$S_k = \frac{1}{\lambda_n - \alpha}.$$
30.2. THE SHIFTED INVERSE POWER METHOD

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 \(u_1 = \sum_{j=1}^m a_j x_j + y\) 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 30.2.2 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,
\[
u_1 = \begin{pmatrix}
1 \\
1 \\
1
\end{pmatrix}
\]

We want the eigenvalue closest to \(-7\). Thus we could use the above method. First we find
\[
\begin{pmatrix}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{pmatrix} + 7 \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix} = \begin{pmatrix}
\frac{68}{133} & \frac{122}{133} & -\frac{65}{133} \\
\frac{4}{133} & \frac{15}{133} & \frac{4}{133} \\
-\frac{3}{7} & -\frac{6}{7} & \frac{4}{7}
\end{pmatrix}
\]

Then beginning with \(u_1\) above, the next iterate is
\[
    u_2 = \begin{pmatrix}
\frac{68}{133} & \frac{122}{133} & -\frac{65}{133} \\
\frac{4}{133} & \frac{15}{133} & \frac{4}{133} \\
-\frac{3}{7} & -\frac{6}{7} & \frac{4}{7}
\end{pmatrix} \begin{pmatrix}
1 \\
1 \\
1
\end{pmatrix} = \begin{pmatrix}
.939849624 \\
.172932331 \\
-.714285714
\end{pmatrix}
\]
Thus $S_2 = 0.939849624$ and

$$
\begin{pmatrix}
0.939849624 \\
0.172932331 \\
-0.714285714
\end{pmatrix}
\frac{1}{0.939849624}
= \begin{pmatrix}
1.0 \\
0.184 \\
-0.76
\end{pmatrix}
$$

Then doing another iteration,

$$
u_3 = \begin{pmatrix}
68 \\
\frac{122}{133} \\
-\frac{3}{7}
\end{pmatrix}
\begin{pmatrix}
1.0 \\
0.184 \\
-0.76
\end{pmatrix}
= \begin{pmatrix}
1.05148872 \\
2.79699248 \times 10^{-2} \\
-1.02057143
\end{pmatrix}
$$

Dividing by the largest element, this yields

$$
\frac{1}{1.05148872}
\begin{pmatrix}
1.05148872 \\
2.79699248 \times 10^{-2} \\
-1.02057143
\end{pmatrix}
= \begin{pmatrix}
1.0 \\
2.66003089 \times 10^{-2} \\
-0.970596651
\end{pmatrix}
$$

The next iteration is

$$
u_4 = \begin{pmatrix}
68 \\
\frac{122}{133} \\
-\frac{3}{7}
\end{pmatrix}
\begin{pmatrix}
1.0 \\
2.66003089 \times 10^{-2} \\
-0.970596651
\end{pmatrix}
= \begin{pmatrix}
1.01003023 \\
3.88434609 \times 10^{-3} \\
-1.00599835
\end{pmatrix}
$$

The scaling factors are not changing by very much so this looks like a good time to stop.

Thus you solve the following for $\lambda$.

$$
\frac{1}{\lambda + 7} = 1.01003023.
$$

This yields $\lambda = 1.01003023$ which yields $\lambda = -6.009930$. This is pretty close to the true eigenvalue, $-6$. How well does $u_4$ work as an eigenvector?

$$
\begin{pmatrix}
5 \\
-4 \\
3
\end{pmatrix}
\begin{pmatrix}
1.01003023 \\
3.88434609 \times 10^{-3} \\
-1.00599835
\end{pmatrix}
= \begin{pmatrix}
-6.07021155 \\
-5.9013564 \times 10^{-4} \\
6.07139182
\end{pmatrix}
$$

while

$$
\begin{pmatrix}
1.01003023 \\
3.88434609 \times 10^{-3} \\
-1.00599835
\end{pmatrix}
\begin{pmatrix}
-6.07021098 \\
-2.33446481 \times 10^{-2} \\
6.04597966
\end{pmatrix}
= \begin{pmatrix}
1.01003023 \\
3.88434609 \times 10^{-3} \\
-1.00599835
\end{pmatrix}
$$

Example 30.2.3 Consider the symmetric matrix, $A = \begin{pmatrix}
1 & 2 & 3 \\
2 & 1 & 4 \\
3 & 4 & 2
\end{pmatrix}$. Find the middle eigenvalue and an eigenvector which goes with it.

Since $A$ is symmetric, it follows it has three real eigenvalues which are solutions to

$$
p(\lambda) = \det \begin{pmatrix}
\lambda & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
- \begin{pmatrix}
1 & 2 & 3 \\
2 & 1 & 4 \\
3 & 4 & 2
\end{pmatrix}
\end{pmatrix}
= \lambda^3 - 4\lambda^2 - 24\lambda - 17 = 0
$$
30.2. **THE SHIFTED INVERSE POWER METHOD**

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

\[
\begin{pmatrix}
-0.8 \\
0 \\
0
\end{pmatrix}
\begin{pmatrix}
1 & 0 & 0 \\
2 & 1 & 4 \\
3 & 4 & 2
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= \begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}
\]

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 \(\alpha = −.855\). You know this is close to the true eigenvalue. Then you find

\[
\begin{pmatrix}
1 & 2 & 3 \\
2 & 1 & 4 \\
3 & 4 & 2
\end{pmatrix}
+.855
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
^{-1}
= \begin{pmatrix}
-367.501105 & 215.95547 & 83.6012034 \\
83.6012034 & -48.7530632 & -19.1913686
\end{pmatrix}
\begin{pmatrix}
1 \\
1 \\
1
\end{pmatrix}
= \begin{pmatrix}
-67.9444316 \\
40.0333028 \\
15.6567716
\end{pmatrix}
\]

The first step of the iteration is then

\[
u_1 = \begin{pmatrix}
-367.501105 & 215.95547 & 83.6012034 \\
83.6012034 & -48.7530632 & -19.1913686
\end{pmatrix}
\begin{pmatrix}
1 \\
1 \\
1
\end{pmatrix}
= \begin{pmatrix}
-67.9444316 \\
40.0333028 \\
15.6567716
\end{pmatrix}
\]

Dividing by the largest entry to normalize the vector on the right,

\[
\begin{pmatrix}
-67.9444316 \\
40.0333028 \\
15.6567716
\end{pmatrix}
\frac{1}{-67.9444316} = \begin{pmatrix}
1.0 \\
-0.58920653 \\
-0.230434949
\end{pmatrix}
\]

Then the next approximation is

\[
u_2 = \begin{pmatrix}
-367.501105 & 215.95547 & 83.6012034 \\
83.6012034 & -48.7530632 & -19.1913686
\end{pmatrix}
\begin{pmatrix}
1.0 \\
-0.58920653 \\
-0.230434949
\end{pmatrix}
= \begin{pmatrix}
-514.008117 \\
302.118746 \\
116.749189
\end{pmatrix}
\]

Divide this by the largest element.

\[
\begin{pmatrix}
-514.008117 \\
302.118746 \\
116.749189
\end{pmatrix}
\frac{1}{-514.008117} = \begin{pmatrix}
1.0 \\
-0.58777038 \\
-0.227134913
\end{pmatrix}
\]
Clearly these vectors are not changing much. An approximate eigenvector is then

\[
\begin{pmatrix}
1.0 \\
-.58777038 \\
-.227134913
\end{pmatrix}
\]

and to find the eigenvalue you solve \( \frac{1}{\lambda + .855} = -514.008117 \), which yields \( \lambda = -.856945495 \).

How well does it work?

\[
\begin{pmatrix}
1 & 2 & 3 \\
2 & 1 & 4 \\
3 & 4 & 2
\end{pmatrix}
\begin{pmatrix}
1.0 \\
-.58777038 \\
-.227134913
\end{pmatrix} = \begin{pmatrix}
-.856945499 \\
.503689968 \\
.194648654
\end{pmatrix}
\]

while

\[
\begin{pmatrix}
-0.856945495 \\
-0.58777038 \\
-0.227134913
\end{pmatrix} = \begin{pmatrix}
-0.856945495 \\
.503687179 \\
.19464224
\end{pmatrix}
\]

I think you can see that for all practical purposes, this has found the eigenvalue and an eigenvector.
Part XI

The Correct Version Of The Riemann Integral *
Appendix A

The Theory Of The Riemann Integral

A.1 An Important Warning

If you read and understand this appendix on the Riemann integral you will become abnormal if you are not already that way. You will laugh at atrocious puns. You will be unpopular with well adjusted confident people. Furthermore, your confidence will be completely shattered. Virtually nothing will be obvious to you ever again. Consider whether it would be better to accept the superficial presentation given earlier than to attempt to acquire deep understanding of the integral, risking your self esteem and confidence, before proceeding further.

A.2 The Definition Of The Riemann Integral

The definition of the Riemann integral of a function of $n$ variables uses the following definition.

Definition A.2.1 For $i = 1, \cdots, n$, let $\{ \alpha^i_k \}_{k=-\infty}^{\infty}$ be points on $\mathbb{R}$ which satisfy

$$\lim_{k \to \infty} \alpha^i_k = \infty, \quad \lim_{k \to -\infty} \alpha^i_k = -\infty, \quad \alpha^i_k < \alpha^i_{k+1}. \quad (1.1)$$

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

$$Q = \prod_{i=1}^{n} [\alpha^i_{j_i}, \alpha^i_{j_i+1}]. \quad (1.2)$$

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 $\mathcal{F}$.
Lemma A.2.2 If \( G \) and \( F \) are two grids, they have a common refinement, denoted here by \( G \lor F \).

**Proof:** Let \( \{ \alpha_k^i \}_{k=-\infty}^{\infty} \) be the sequences used to construct \( G \) and let \( \{ \beta_k^i \}_{k=-\infty}^{\infty} \) be the sequence used to construct \( F \). Now let \( \{ \gamma_k^i \}_{k=-\infty}^{\infty} \) denote the union of \( \{ \alpha_k^i \}_{k=-\infty}^{\infty} \) and \( \{ \beta_k^i \}_{k=-\infty}^{\infty} \). It is necessary to show that for each \( i \) these points can be arranged in order. To do so, let \( \gamma_0^i \equiv \alpha_0^i \). Now if \( \gamma_{-j}^i, \ldots, \gamma_0^i, \ldots, \gamma_j^i \) have been chosen such that they are in order and all distinct, let \( \gamma_{j+1}^i \) be the first element of
\[
\{ \alpha_k^i \}_{k=-\infty}^{\infty} \cup \{ \beta_k^i \}_{k=-\infty}^{\infty}
\] (1.3) which is larger than \( \gamma_j^i \) and let \( \gamma_{-(j+1)}^i \) be the last element of \( \{ \gamma_k^i \}_{-\infty}^{\infty} \) which is strictly smaller than \( \gamma_{-j}^i \). The assumption \( \{ \alpha_k^i \}_{-\infty}^{\infty} \) insures such a first and last element exists. Now let the grid \( G \lor F \) consist of boxes of the form
\[
Q \equiv \prod_{i=1}^{n} [\gamma_j^i, \gamma_{j+1}^i].
\]

The Riemann integral is only defined for functions, \( f \) which are bounded and are equal to zero off some bounded set, \( D \). In what follows \( f \) will always be such a function.

**Definition A.2.3** Let \( f \) be a bounded function which equals zero off a bounded set, \( D \), and let \( G \) be a grid. For \( Q \in G \), define
\[
M_Q(f) \equiv \sup \{ f(x) : x \in Q \}, \quad m_Q(f) \equiv \inf \{ f(x) : x \in Q \}.
\] (1.4)

Also define for \( Q \) a box, the volume of \( Q \), denoted by \( v(Q) \) by
\[
v(Q) = \prod_{i=1}^{n} (b_i - a_i), \quad Q \equiv \prod_{i=1}^{n} [a_i, b_i].
\]

Now define upper sums, \( U_G(f) \) and lower sums, \( L_G(f) \) with respect to the indicated grid, by the formulas
\[
U_G(f) \equiv \sum_{Q \in G} M_Q(f) v(Q), \quad L_G(f) \equiv \sum_{Q \in G} m_Q(f) v(Q).
\]

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 \( Q \in G \) so there are no convergence questions to be considered here.

**Lemma A.2.4** If \( F \) is a refinement of \( G \) then
\[
U_G(f) \geq U_F(f), \quad L_G(f) \leq L_F(f).
\]

Also if \( F \) and \( G \) are two grids,
\[
L_G(f) \leq U_F(f).
\]
A.2. THE DEFINITION OF THE RIEMANN INTEGRAL

**Proof:** For $P \in \mathcal{G}$ let $\tilde{P}$ denote the set,

\[ \{Q \in \mathcal{F} : Q \subseteq P\}. \]

Then $P = \cup \tilde{P}$ and

\[
\mathcal{L}_F(f) \equiv \sum_{Q \in \mathcal{F}} m_Q(f) v(Q) = \sum_{P \in \mathcal{G}} \sum_{Q \in \tilde{P}} m_Q(f) v(Q) \\
\geq \sum_{P \in \mathcal{G}} m_P(f) \sum_{Q \in \tilde{P}} v(Q) = \sum_{P \in \mathcal{G}} m_P(f) v(P) \equiv \mathcal{L}_G(f).
\]

Similarly, the other inequality for the upper sums is valid.

To verify the last assertion of the lemma, use Lemma A.2.2 to write

\[
\mathcal{L}_G(f) \leq \mathcal{L}_{G \vee F}(f) \leq \mathcal{U}_{G \vee F}(f) \leq \mathcal{U}_F(f).
\]

This proves the lemma.

This lemma makes it possible to define the Riemann integral.

**Definition A.2.5** Define an upper and a lower integral as follows.

\[
\mathcal{T}(f) \equiv \inf \{\mathcal{U}_G(f) : G \text{ is a grid}\},
\]

\[
\mathcal{I}(f) \equiv \sup \{\mathcal{L}_G(f) : G \text{ is a grid}\}.
\]

**Lemma A.2.6** $\mathcal{T}(f) \geq \mathcal{I}(f)$.

**Proof:** From Lemma A.2.3 it follows for any two grids $G$ and $F$,

\[
\mathcal{L}_G(f) \leq \mathcal{U}_F(f).
\]

Therefore, taking the supremum for all grids on the left in this inequality,

\[
\mathcal{I}(f) \leq \mathcal{U}_F(f)
\]

for all grids $F$. Taking the infimum in this inequality, yields the conclusion of the lemma.

**Definition A.2.7** A bounded function, $f$ which equals zero off a bounded set, $D$, is said to be Riemann integrable, written as $f \in \mathcal{R}(\mathbb{R}^n)$ exactly when $\mathcal{I}(f) = \mathcal{T}(f)$. In this case define

\[
\int f \, dV \equiv \int f \, dx = \mathcal{T}(f) = \mathcal{I}(f).
\]

As in the case of integration of functions of one variable, one obtains the Riemann criterion which is stated as the following theorem.

**Theorem A.2.8** (Riemann criterion) $f \in \mathcal{R}(\mathbb{R}^n)$ if and only if for all $\varepsilon > 0$ there exists a grid $G$ such that

\[
\mathcal{U}_G(f) - \mathcal{L}_G(f) < \varepsilon.
\]
Proof: If \( f \in \mathcal{R} (\mathbb{R}^n) \), then \( T(f) = I(f) \) and so there exist grids \( \mathcal{G} \) and \( \mathcal{F} \) such that
\[
U_{\mathcal{G}}(f) - L_{\mathcal{F}}(f) \leq T(f) + \frac{\varepsilon}{2} - \left(I(f) - \frac{\varepsilon}{2}\right) = \varepsilon.
\]
Then letting \( \mathcal{H} = \mathcal{G} \lor \mathcal{F} \), Lemma A.2.4 implies
\[
U_{\mathcal{H}}(f) - L_{\mathcal{H}}(f) \leq U_{\mathcal{G}}(f) - L_{\mathcal{F}}(f) < \varepsilon.
\]
Conversely, if for all \( \varepsilon > 0 \) there exists \( \mathcal{G} \) such that
\[
U_{\mathcal{G}}(f) - L_{\mathcal{G}}(f) < \varepsilon,
\]
then
\[
T(f) - I(f) \leq U_{\mathcal{G}}(f) - L_{\mathcal{G}}(f) < \varepsilon.
\]
Since \( \varepsilon > 0 \) is arbitrary, this proves the theorem.

A.3 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 A.3.1 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
\[
h (x) = \phi (f (x), g (x)),
\]
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 \( k = f, g \),
\[
U_{\mathcal{G}}(k) - L_{\mathcal{G}}(k) < \delta^2. \quad (1.5)
\]
Then defining for \( k = f, g \),
\[
\mathcal{P}_k \equiv \{ Q \in \mathcal{G} : M_Q (k) - m_Q (k) > \delta \},
\]
it follows
\[
\delta^2 > \sum_{Q \in \mathcal{G}} (M_Q (k) - m_Q (k)) v (Q) \geq \sum_{\mathcal{P}_k} (M_Q (k) - m_Q (k)) v (Q) \geq \delta \sum_{\mathcal{P}_k} v (Q)
\]
and so for \( k = f, g \),
\[
\varepsilon > \delta > \sum_{\mathcal{P}_k} v (Q). \quad (1.6)
\]
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
\[
S \equiv \{Q \in \mathcal{G} : 0 < M_Q(k) - m_Q(k) \leq \delta, k = f, g\}.
\]
Thus the union of the boxes in \(S\) 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
\[
U_G(h) - L_G(h) \leq \sum_{Q \in P_I} (M_Q(h) - m_Q(h))v(Q) + \sum_{Q \in P_a} (M_Q(h) - m_Q(h))v(Q) + \sum_{Q \in S} \delta v(Q).
\]
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
\[
U_G(h) - L_G(h) \leq C \sum_{Q \in P_I} v(Q) + C \sum_{Q \in P_a} v(Q) + \sum_{Q \in S} \delta v(Q),
\]
which by (1.8) 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 A.3.2** 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
\[
\int |f| \, dx \geq \left| \int f \, dx \right|. \tag{1.7}
\]

**Proof:** Each of the combinations of functions described above is Riemann integrable by Theorem A.3.1. 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 \(\mathcal{G}\) be a grid with the property that for \(k = f, g, |f|\) and \(af + bg\),
\[
U_G(k) - L_G(k) < \varepsilon. \tag{1.9}
\]
Consider (1.9). For each \(Q \in \mathcal{G}\) pick a point in \(Q, x_Q\). Then
\[
\sum_{Q \in \mathcal{G}} k(x_Q)v(Q) \in [L_G(k), U_G(k)].
\]
and so
\[ \left| \int k \, dx - \sum_{Q \in \mathcal{G}} k(\mathbf{x}_Q) \, v(Q) \right| < \varepsilon. \]

Consequently, since
\[ \sum_{Q \in \mathcal{G}} (af + bg)(\mathbf{x}_Q) \, v(Q) = a \sum_{Q \in \mathcal{G}} f(\mathbf{x}_Q) \, v(Q) + b \sum_{Q \in \mathcal{G}} g(\mathbf{x}_Q) \, v(Q), \]
it follows
\[ \left| \int (af + bg) \, dx - a \int f \, dx - b \int g \, dx \right| \leq \left| \int (af + bg) \, dx - \sum_{Q \in \mathcal{G}} (af + bg)(\mathbf{x}_Q) \, v(Q) \right| + \]
\[ \left| a \sum_{Q \in \mathcal{G}} f(\mathbf{x}_Q) \, v(Q) - a \int f \, dx \right| + \left| b \sum_{Q \in \mathcal{G}} g(\mathbf{x}_Q) \, v(Q) - b \int g \, dx \right| \]
\[ \leq \varepsilon + |a| \varepsilon + |b| \varepsilon. \]

Since \( \varepsilon \) is arbitrary, this establishes Formula 1.7 and shows the integral is linear.

It remains to establish the inequality 1.8. By 1.9, and the triangle inequality for sums,
\[ \int |f| \, dx + \varepsilon \geq \sum_{Q \in \mathcal{G}} |f(\mathbf{x}_Q)| \, v(Q) \geq \]
\[ \geq \left| \sum_{Q \in \mathcal{G}} f(\mathbf{x}_Q) \, v(Q) \right| \geq \left| \int f \, dx \right| - \varepsilon. \]

Then since \( \varepsilon \) is arbitrary, this establishes the desired inequality. This proves the corollary.

### A.4 Which Functions Are Integrable?

Which functions are in \( \mathcal{R}(\mathbb{R}^n) \)? As in the case of integrals of functions of one variable, this is an important question. It turns out the Riemann integrable functions are characterized by being continuous except on a very small set. This has to do with Jordan content.

**Definition A.4.1** A bounded set, \( E \), has Jordan content 0 or content 0 if for every \( \varepsilon > 0 \) there exists a grid, \( \mathcal{G} \) such that
\[ \sum_{Q \cap E \neq \emptyset} v(Q) < \varepsilon. \]

This symbol says to sum the volumes of all boxes from \( \mathcal{G} \) which have nonempty intersection with \( E \).

Next it is necessary to define the oscillation of a function.
Definition A.4.2 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 A.4.3 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 A.4.4 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 A.4.5 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 A.4.6 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
\[
\sum_{Q \in \mathcal{G}_W} v(Q) < \varepsilon \tag{1.10}
\]
where
\[
\mathcal{G}_W \equiv \{Q \in \mathcal{G} : Q \cap W \neq \emptyset\}.
\]

Proof: Let \( W \) have content zero. Also let \( |f(x)| < C/2 \) for all \( x \in \mathbb{R}^n \), let \( \varepsilon > 0 \) be given, and let \( \mathcal{G} \) be a grid which satisfies (1.10). Since \( f \) equals zero off some bounded set, there exists \( R \) such that \( f \) equals zero off of \( B(0,\frac{C}{2}) \). Thus \( W \subseteq B(0,\frac{C}{2}) \). Also note that if \( \mathcal{G} \) is a grid for which (1.10) holds, then this inequality continues to hold if \( \mathcal{G} \) is replaced with a refined grid. Therefore, you may assume the diameter of every box in \( \mathcal{G} \) which intersects
B(0, R) is less than $R/3$ and so all boxes of $G$ which intersect the set where $f$ is nonzero are contained in $B(0, R)$. 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.

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

$$\prod_{i=1}^{n} (b_i + \eta_i - (a_i - \eta_i)) = v(\tilde{Q}),$$

and $G_W$ denotes those $\tilde{Q}$ for $Q \in G$ which have nonempty intersection with $W$, then

$$\sum_{\tilde{Q} \in \tilde{G}_W} v(\tilde{Q}) < \varepsilon$$

(1.11)

where $\tilde{Q}$ is the box,

$$\prod_{i=1}^{n} (b_i + 2\eta_i - (a_i - 2\eta_i))$$

For each $x \in \mathbb{R}^n$, let $r_x < \min(\eta_1/2, \cdots, \eta_n/2)$ be such that

$$\omega_{f, r_x}(x) < \varepsilon + \omega_f(x).$$

(1.12)

Now let $C$ denote all intersections of the form $\tilde{Q} \cap B(x, r_x)$ such that $x \in \overline{B(0, R)}$ so that $C$ is an open cover of the compact set, $\overline{B(0, R)}$. Let $\delta$ be a Lebesgue number for this open cover of $\overline{B(0, R)}$ 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 $C$. Let $C_W$ be those open sets of $C$, $\tilde{Q} \cap B(x, r_x)$, for which $x \in W$. Thus each of these sets is contained in some $\tilde{Q}$ where $Q \in G_W$. Let $F_W$ be those sets of $F_1$ which are subsets of some set of $C_W$. Thus

$$\sum_{Q \in F_W} v(Q) < \varepsilon.$$ 

(1.13)

because each $Q$ in $F_W$ is contained in a set, $\tilde{Q}$ described above and the sum of the volumes of these is less than $\varepsilon$. Then

$$U_f(f) - L_f(f) = \sum_{Q \in F_W} (M_Q(f) - m_Q(f)) v(Q)$$
A.4. WHICH FUNCTIONS ARE INTEGRABLE?

\[ + \sum_{Q \in \mathcal{F} \setminus \mathcal{F}_W} (M_Q(f) - m_Q(f)) v(Q). \]

If \( Q \in \mathcal{F} \setminus \mathcal{F}_W \), then \( Q \) must be a subset of some set of \( \mathcal{C} \setminus \mathcal{C}_W \) since it is not in any set of \( \mathcal{C}_W \). Say \( Q \subseteq Q_1 \cap B(x, r_x) \) where \( x \notin W \). Therefore, from 1.12 and the observation that \( x \notin W \), it follows \( \omega_f(x) = 0 \) and so

\[ M_Q(f) - m_Q(f) \leq \varepsilon. \]

Therefore, from 1.13 and the estimate on \( f \),

\[ \mathcal{U}_F(f) - \mathcal{L}_F(f) \leq \sum_{Q \in \mathcal{F}_W} C v(Q) + \sum_{Q \in \mathcal{F} \setminus \mathcal{F}_W} \varepsilon v(Q) \leq C \varepsilon + \varepsilon (2R)^n, \]

the estimate of the second sum coming from the fact

\[ B(0, R) \subseteq \prod_{i=1}^n [-R, R]. \]

Since \( \varepsilon \) is arbitrary, this proves the theorem.

**Definition A.4.7** A bounded set, \( E \) is a Jordan set in \( \mathbb{R}^n \) or a contented set in \( \mathbb{R}^n \) if

\[ \mathcal{X}_E(x) = \begin{cases} 
1 & \text{if } x \in E \\
0 & \text{if } x \notin E
\end{cases} \]

It is called the indicator function because it indicates whether \( x \) is in \( E \) according to whether it equals 1. For a function \( f \in \mathcal{R} (\mathbb{R}^n) \) and \( E \) a contented set, \( f \mathcal{X}_E \in \mathcal{R} (\mathbb{R}^n) \) by Corollary A.3.2. Then

\[ \int_E f dV \equiv \int f \mathcal{X}_E dV. \]

So what are examples of contented sets?

**Theorem A.4.8** Suppose \( E \) is a bounded contented set in \( \mathbb{R}^n \) and \( f, g : E \rightarrow \mathbb{R} \) are two functions satisfying \( f(x) \geq g(x) \) for all \( x \in E \) and \( f \mathcal{X}_E \) and \( g \mathcal{X}_E \) 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 \( k = f \mathcal{X}_E, g \mathcal{X}_E \),

\[ \mathcal{U}_G(k) - \mathcal{L}_G(k) < \varepsilon / 4. \quad (1.14) \]

\(^*\)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 \( \varepsilon > 0, \text{Lemma} \) 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.
Also let \( K \geq \sum_{j=1}^{m} v_n(Q_j) \) where the \( Q_j \) are the boxes which intersect \( E \). Let \( \{a_i\}_{i=-\infty}^{\infty} \) be a sequence on \( \mathbb{R} \), \( a_i < a_{i+1} \) for all \( i \), which includes

\[
M_{Q_j}(f \chi_E) + \frac{\varepsilon}{4mK}, M_{Q_j}(f \chi_E), M_{Q_j}(g \chi_E), \]

\[
m_{Q_j}(f \chi_E), m_{Q_j}(g \chi_E), m_{Q_j}(g \chi_E) - \frac{\varepsilon}{4mK}
\]

for all \( j = 1, \cdots, m \). Now define a grid on \( \mathbb{R}^{n+1} \) as follows.

\[
G' \equiv \{ Q \times [a_i, a_{i+1}] : Q \in G, \ i \in \mathbb{Z} \}
\]

In words, this grid consists of all possible boxes of the form \( Q \times [a_i, a_{i+1}] \) where \( Q \in G \) and \( a_i \) is a term of the sequence just described. It is necessary to verify that for \( P \in G' \), \( \chi_P \in \mathcal{R}(\mathbb{R}^{n+1}) \). This is done by showing that \( \mathcal{U}_{G'}(\chi_P) - \mathcal{L}_{G'}(\chi_P) < \varepsilon \) and then noting that \( \varepsilon > 0 \) was arbitrary. For \( G' \) just described, denote by \( Q' \) a box in \( G' \). Thus \( Q' = Q \times [a_i, a_{i+1}] \) for some \( i \).

\[
\mathcal{U}_{G'}(\chi_P) - \mathcal{L}_{G'}(\chi_P) = \sum_{Q' \in G'} (M_{Q'}(\chi_P) - m_{Q'}(\chi_P)) v_{n+1}(Q')
\]

\[
= \sum_{i=-\infty}^{\infty} \sum_{j=1}^{m} (M_{Q_j}(\chi_P) - m_{Q_j}(\chi_P)) v_n(Q_j)(a_{i+1} - a_i)
\]

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

\[
\mathcal{U}_{G'}(\chi_P) - \mathcal{L}_{G'}(\chi_P) = \sum_{j=1}^{m} v_n(Q_j) \sum_{i=-\infty}^{\infty} (M_{Q_j \times [a_i, a_{i+1}]}(\chi_P) - m_{Q_j \times [a_i, a_{i+1}]}(\chi_P))(a_{i+1} - a_i).
\]

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

\[
M_{Q_j \times [a_i, a_{i+1}]}(\chi_P) - m_{Q_j \times [a_i, a_{i+1}]}(\chi_P) = 0.
\]

You see, \( \chi_P \) 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,

\[
M_{Q_j \times [a_i, a_{i+1}]}(\chi_P) - m_{Q_j \times [a_i, a_{i+1}]}(\chi_P) = 0
\]
on these boxes. However, on the boxes intersected by the surfaces, the value of
\[ M_{Q_j \times [a_i, a_{i+1}]}(X_P) - m_{Q_j \times [a_i, a_{i+1}]}(X_P) \]
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
\[ M_{Q_j}(fX_E) + \frac{\varepsilon}{4mK}, M_{Q_j}(fX_E), M_{Q_j}(gX_E), m_{Q_j}(fX_E), m_{Q_j}(gX_E) \]
for all \( j = 1, \ldots, m \),
\[ \sum_{i=-\infty}^{\infty} (M_{Q_j \times [a_i, a_{i+1}]}(X_P) - m_{Q_j \times [a_i, a_{i+1}]}(X_P)) (a_{i+1} - a_i) \leq \]
\[ \sum_{\{i: m_{Q_j}(gX_E) \leq a_i < M_{Q_j}(gX_E)\}} 1(a_{i+1} - a_i) + \sum_{\{i: m_{Q_j}(fX_E) \leq a_i < M_{Q_j}(fX_E)\}} 1(a_{i+1} - a_i) \quad (1.15) \]
The first of the sums in \( U_P \) contains all possible terms for which
\[ M_{Q_j \times [a_i, a_{i+1}]}(X_P) - m_{Q_j \times [a_i, a_{i+1}]}(X_P) \]
might be 1 due to the graph of the bottom surface \( gX_E \) while the second sum contains all possible terms for which the expression might be 1 due to the graph of the top surface \( fX_E \).
\[ \leq \left( M_{Q_j}(gX_E) + \frac{\varepsilon}{4mK} - m_{Q_j}(gX_E) \right) + \left( M_{Q_j}(fX_E) + \frac{\varepsilon}{4mK} - m_{Q_j}(fX_E) \right) \]
\[ = (M_{Q_j}(gX_E) - m_{Q_j}(gX_E)) + (M_{Q_j}(fX_E) - m_{Q_j}(fX_E)) + \frac{\varepsilon}{2m} \left( \sum_{j=1}^{m} v(Q_j) \right)^{-1} \]
Therefore, by \( U_P \)
\[ U_{Q'}(X_P) - L_{Q'}(X_P) \leq \]
\[ \sum_{j=1}^{m} v(Q_j) \left[ (M_{Q_j}(gX_E) - m_{Q_j}(gX_E)) + (M_{Q_j}(fX_E) - m_{Q_j}(fX_E)) \right] \]
\[ + \frac{m}{2m} \left( \sum_{j=1}^{m} v(Q_j) \right)^{-1} \]
\[ = U_{\mathcal{G}}(f) - L_{\mathcal{G}}(f) + U_{\mathcal{G}}(g) - L_{\mathcal{G}}(g) + \frac{\varepsilon}{2} \]
\[ < \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{2} = \varepsilon. \]
Since \( \varepsilon > 0 \) is arbitrary, this proves the theorem.

**Corollary A.4.9** 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: Since $E$ is contented, meaning $X_E$ is integrable, it follows from Theorem A.4.6 the set of discontinuities of $X_E$ has Jordan content 0. But the set of discontinuities of $X_E$ is $\partial E$ defined as those points $x$ such that $B(x, r)$ contains points of $E$ and points of $E^C$ for every $r > 0$. Extend $f$ and $g$ to equal 0 off $E$. Then the set of discontinuities of these extended functions, still denoted as $f, g$ is $\partial E$ which has Jordan content 0. This reduces to the situation of Theorem A.4.8. 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, it follows that every box is a contented set. Therefore, if $B_i$ is a box, functions of the form

$$\sum_{i=1}^{m} a_i X_{B_i}$$

are integrable. These functions are called step functions.

The following theorem is analogous to the fact that in one dimension, when you integrate over a point, the answer is 0.

Theorem A.4.10 If a bounded set, $E$, has Jordan content 0, then $E$ is a Jordan (contented) set and if $f$ is any bounded function defined on $E$, then $fX_E \in \mathcal{R}(\mathbb{R}^n)$ and

$$\int_E f \, dV = 0.$$

Proof: Let $G$ be a grid with

$$\sum_{Q \cap E \neq \emptyset} v(Q) < \frac{\varepsilon}{1 + (M - m)}.$$

Then

$$U_G(fX_E) \leq \sum_{Q \cap E \neq \emptyset} Mv(Q) \leq \frac{\varepsilon M}{1 + (M - m)}$$

and

$$L_G(fX_E) \geq \sum_{Q \cap E \neq \emptyset} mv(Q) \geq \frac{\varepsilon m}{1 + (M - m)}$$

and so

$$U_G(fX_E) - L_G(fX_E) \leq \sum_{Q \cap E \neq \emptyset} Mv(Q) - \sum_{Q \cap E \neq \emptyset} mv(Q)$$

$$= (M - m) \sum_{Q \cap E \neq \emptyset} v(Q) < \frac{\varepsilon (M - m)}{1 + (M - m)} < \varepsilon.$$
A.4. WHICH FUNCTIONS ARE INTEGRABLE?

This shows \( f_X \in \mathcal{R}(\mathbb{R}^n) \). Now also,
\[
m \varepsilon \leq \int f_X \, dV \leq M \varepsilon
\]
and since \( \varepsilon \) is arbitrary, this shows
\[
\int_E f \, dV = \int f_X \, dV = 0
\]
Why is \( E \) contented? Let \( \mathcal{G} \) be a grid for which
\[
\sum_{Q \cap E \neq \emptyset} v(Q) < \varepsilon
\]
Then for this grid,
\[
\mathcal{U}_G(X_E) - \mathcal{L}_G(X_E) \leq \sum_{Q \cap E \neq \emptyset} v(Q) < \varepsilon
\]
and this proves the theorem.

**Corollary A.4.11** If \( f_X \) \( i \in \mathcal{R}(\mathbb{R}^n) \) 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 \( f_X \in \mathcal{R}(\mathbb{R}^n) \) and
\[
\int f_X \, dV = \sum_{i=1}^r \int_{E_i} f \, dV.
\]

**Proof:** 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, \( f_X \) \( i \in \mathcal{R}(\mathbb{R}^n) \). 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.

\[
-f_{X_{F\cap E_{r+1}}} + f_{X_{F_r}} + f_{X_{E_{r+1}}} = f_{X_F}
\]
and by Theorem A.4.10 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,
\[
\int f_{X_{F_{r+1}}} \, dV = \int f_{X_{F_r}} \, dV + \int f_{X_{E_{r+1}}} \, dV
\]
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.

In particular, for
\[
Q = \prod_{i=1}^n [a_i, b_i], \quad Q' = \prod_{i=1}^n (a_i, b_i)
\]
both are contented sets and
\[
\int X_Q \, dV = \int_{Q'} X_{Q'} \, dV = v(Q).
\]
(1.16)
This is because
\[ Q \setminus Q' = \bigcup_{i=1}^{n} a_i \times \prod_{j \neq i} (a_j, b_j) \]
a finite union of sets of content 0. It is obvious \( \int x_Q \, dV = v(Q) \) because you can use a grid which has \( Q \) as one of the boxes and then the upper and lower sums are the same and equal to \( v(Q) \). Therefore, the claim about the equality of the two integrals in \( \text{[4.10]} \) follows right away from Corollary \( \text{[4.11]} \). That \( \chi_{Q'} \) is integrable follows from
\[ \chi_{Q'} = \chi_Q - \chi_{Q \setminus Q'} \]
and each of the two functions on the right is integrable thanks to Theorem \( \text{[4.11]} \).

In fact, here is an interesting version of the Riemann criterion which depends on these half open boxes.

**Lemma A.4.12** 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
\[ \sum_{Q \in \mathcal{G}} (M_Q (f) - m_Q (f)) \, v(Q) < \varepsilon. \tag{1.17} \]

**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 \( \text{[4.10]} \) holds with \( \varepsilon \) replaced with \( \frac{\varepsilon}{2} \). It is necessary to show there is a grid such that \( \text{[4.10]} \) holds with no primes on the \( Q \). Let \( \mathcal{F} \) be a refinement of \( \mathcal{G} \) obtained by adding the points \( \alpha_k + \eta_k \) where \( \eta_k \leq \eta \) and is also chosen so small that for each \( i = 1, \ldots, n \),
\[ \alpha_k^i + \eta_k < \alpha_{k+1}^i. \]
You only need to have \( \eta_k > 0 \) for the finitely many boxes of \( \mathcal{G} \) which intersect the bounded set where \( f \) is not zero. Then for
\[ Q = \prod_{i=1}^{n} [\alpha_{k_i}, \alpha_{k_i+1}] \in \mathcal{G}, \]
Let
\[ \mathcal{\hat{G}} = \prod_{i=1}^{n} [\alpha_{k_i} + \eta_k, \alpha_{k_i+1}] \]
and denote by \( \mathcal{\hat{G}} \) the collection of these smaller boxes. For each set, \( Q \) in \( \mathcal{G} \) there is the smaller set, \( \mathcal{\hat{Q}} \) 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, \( \mathcal{\hat{Q}} \) is the union of \( \mathcal{\hat{Q}} \) and these sets, \( B_k \). Now suppose \( f \) equals zero off the ball \( B(0, \frac{\varepsilon}{2}) \).

Then without loss of generality, you may assume the diameter of every box in \( \mathcal{\hat{G}} \) which has nonempty intersection with \( B(0, R) \) is smaller than \( \frac{\varepsilon}{2} \). (If this is not so, simply refine \( \mathcal{G} \) to make it so, such a refinement leaving \( \text{[4.10]} \) valid because refinements do not increase the difference between upper and lower sums in this context either.) Suppose there are \( P \) sets of \( \mathcal{G} \) contained in \( B(0, R) \) (So these are the only sets of \( \mathcal{G} \) which could have nonempty intersection with the set where \( f \) is nonzero.) and suppose that for all \( x \), \( |f(x)| < C/2 \). Then
\[ \sum_{Q \in \mathcal{F}} (M_Q (f) - m_Q (f)) \, v(Q) \leq \sum_{Q \in \mathcal{\hat{G}}} (M_{\mathcal{\hat{Q}}} (f) - m_{\mathcal{\hat{Q}}} (f)) \, v(Q) \]
\[ + \sum_{Q \in \mathcal{F} \setminus \mathcal{G}} (M_Q(f) - m_Q(f)) v(Q) \]

The first term on the right of the inequality in the above is no larger than \( \varepsilon / 2 \) because \( M_Q(f) - m_Q(f) \leq M_Q'(f) - m_Q'(f) \) for each \( Q \). Therefore, the above is dominated by

\[ \leq \varepsilon / 2 + CPnR^{n-1}\eta < \varepsilon \]

whenever \( \eta \) is small enough. Since \( \varepsilon \) is arbitrary, \( f \in \mathcal{R}(\mathbb{R}^n) \) as claimed.

### A.5 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 \( \mathbb{R}^{n+m} \),

\[ y \rightarrow f(x, y) \]

is a function of \( y \) for each \( x \in \mathbb{R}^n \). 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 A.5.1** Let \( \mathcal{G} \) be a grid on \( \mathbb{R}^{n+m} \) defined by the \( n+m \) sequences,

\[ \{ \alpha^i_k \}_{k=-\infty}^{\infty} \quad i = 1, \cdots, n+m. \]

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

\[ \prod_{i=n+1}^{n+m} [\alpha^i_{k_i}, \alpha^i_{k_{i+1}}], \quad k_i \geq n + 1 \]

and a box in \( \mathcal{G}_n \) would be of the form

\[ \prod_{i=1}^{n} [\alpha^i_{k_i}, \alpha^i_{k_{i+1}}], \quad k_i \leq n. \]

**Lemma A.5.2** Let \( \mathcal{G}, \mathcal{G}_n, \) and \( \mathcal{G}_m \) be the grids defined above. Then

\[ \mathcal{G} = \{ R \times P : R \in \mathcal{G}_n \text{ and } P \in \mathcal{G}_m \}. \]
APPENDIX A. THE THEORY OF THE RIEMANN INTEGRAL

Proof: If $Q \in \mathcal{G}$, 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 $\mathcal{G}$. This proves the lemma.

Now let $\mathcal{G}$ be a grid on $\mathbb{R}^{n+m}$ and suppose

$$
\phi(z) = \sum_{Q \in \mathcal{G}} \phi_Q \chi_{Q'}(z)
$$

(1.18)

where $\phi_Q$ equals zero for all but finitely many $Q$. Thus $\phi$ is a step function. Recall that for $Q = \prod_{i=1}^{n+m} [a_i, b_i]$, $Q' = \prod_{i=1}^{n+m} (a_i, b_i)$.

The function

$$
\phi = \sum_{Q \in \mathcal{G}} \phi_Q \chi_{Q'}
$$

is integrable because it is a finite sum of integrable functions, each function in the sum being integrable because the set of discontinuities has Jordan content 0. (why?) Letting $(x, y) = z$,

$$
\phi(x, y) = \phi(x, \cdot) = \sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} \phi_{R \times P} \chi_{R' \times P'}(x, \cdot)
$$

(1.19)

For a function of two variables, $h$, denote by $h(\cdot, y)$ the function, $x \rightarrow h(x, y)$ and $h(x, \cdot)$ the function $y \rightarrow h(x, y)$. The following lemma is a preliminary version of Fubini’s theorem.

Lemma A.5.3 Let $\phi$ be a step function as described in 1.18. Then

$$
\int_{\mathbb{R}^m} \phi(\cdot, y) \, dV_y \in \mathcal{R}(\mathbb{R}^n),
$$

(1.20)

and

$$
\int_{\mathbb{R}^n} \int_{\mathbb{R}^m} \phi(x, y) \, dV_y \, dV_x = \int_{\mathbb{R}^{n+m}} \phi(z) \, dV.
$$

(1.22)

Proof: To verify 1.20, note that $\phi(x, \cdot)$ is the step function

$$
\phi(x, \cdot) = \sum_{P \in \mathcal{G}_m} \phi_{R \times P} \chi_{P'}(y).
$$

Where $x \in R'$ and this is a finite sum of integrable functions because each has set of discontinuities with Jordan content 0. From the description in 1.18,

$$
\int_{\mathbb{R}^m} \phi(x, y) \, dV_y = \sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} \phi_{R \times P} \chi_{R'}(x) \, v(P)
$$

(1.23)
Another step function. Therefore,
\[
\int_{\mathbb{R}^n} \int_{\mathbb{R}^m} \phi(x, y) \, dV_y \, dV_x = \sum_{R \in G_n} \sum_{P \in G_m} \phi_{R \times P} \psi(P) \, v(R)
\]
\[
= \sum_{Q \in \tilde{G}} \phi_Q \psi(Q) = \int_{\mathbb{R}^{n+m}} \phi(z) \, dV.
\]
and this proves the lemma.

From (1.23),
\[
M_{R_1'} \left( \int_{\mathbb{R}^m} \phi(\cdot, y) \, dV_y \right) \equiv \sup \left\{ \sum_{R \in G_n} \left( \sum_{P \in G_m} \phi_{R \times P} \psi(P) \right) \chi_{R'}(x) : x \in R_1' \right\}
\]
\[
= \sum_{P \in \tilde{G}_m} \phi_{R_1 \times P} \psi(P) \quad (1.24)
\]
because \( \int_{\mathbb{R}^m} \phi(\cdot, y) \, dV_y \) has the constant value given in (1.23) for \( x \in R_1' \). Similarly,
\[
m_{R_1'} \left( \int_{\mathbb{R}^m} \phi(\cdot, y) \, dV_y \right) \equiv \inf \left\{ \sum_{R \in G_n} \left( \sum_{P \in G_m} \phi_{R \times P} \psi(P) \right) \chi_{R'}(x) : x \in R_1' \right\}
\]
\[
= \sum_{P \in \tilde{G}_m} \phi_{R_1 \times P} \psi(P). \quad (1.25)
\]

**Theorem A.5.4** (Fubini) Let \( f \in \mathcal{R}(\mathbb{R}^{n+m}) \) and suppose also that \( f(x, \cdot) \in \mathcal{R}(\mathbb{R}^m) \) for each \( x \). Then
\[
\int_{\mathbb{R}^m} f(\cdot, y) \, dV_y \in \mathcal{R}(\mathbb{R}^n)
\]
and
\[
\int_{\mathbb{R}^{n+m}} f(z) \, dV = \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} f(x, y) \, dV_y \, dV_x.
\]

**Proof:** Let \( G \) be a grid such that \( U_G(f) - L_G(f) < \varepsilon \) and let \( G_n \) and \( G_m \) be as defined above. Let
\[
\phi(z) \equiv \sum_{Q \in \tilde{G}} M_{Q'}(f) \chi_{Q'}(z), \quad \psi(z) \equiv \sum_{Q \in \tilde{G}} m_{Q'}(f) \chi_{Q'}(z).
\]
Observe that \( M_{Q'}(f) \leq M_Q(f) \) and \( m_{Q'}(f) \geq m_Q(f) \). Then
\[
U_G(f) \geq \int \phi \, dV, \quad L_G(f) \leq \int \psi \, dV.
\]
Also \( f(z) \in (\psi(z), \phi(z)) \) for all \( z \). Thus from (1.24),
\[
M_{R'} \left( \int_{\mathbb{R}^m} f(\cdot, y) \, dV_y \right) \leq M_{R'} \left( \int_{\mathbb{R}^m} \phi(\cdot, y) \, dV_y \right) = \sum_{P \in \tilde{G}_m} M_{R' \times P'}(f) \psi(P)
\]
and from (1.27),
\[
m_{R'} \left( \int_{\mathbb{R}^m} f(\cdot, y) \, dV_y \right) \geq m_{R'} \left( \int_{\mathbb{R}^m} \psi(\cdot, y) \, dV_y \right) = \sum_{P \in \tilde{G}_m} m_{R' \times P'}(f) \psi(P).
\]
Therefore,
\[
\sum_{R \in \mathcal{G}, n} \left[ M_{R'} \left( \int_{\mathbb{R}^m} f (\cdot, y) \, dV_y \right) - m_{R'} \left( \int_{\mathbb{R}^m} f (\cdot, y) \, dV_y \right) \right] v (R) \leq \\
\sum_{R \in \mathcal{G}, n} \sum_{P \in \mathcal{P}} \left[ M_{R \times P'} (f) - m_{R \times P'} (f) \right] v (P) v (R) \leq U_{\mathcal{G}} (f) - L_{\mathcal{G}} (f) < \varepsilon.
\]

This shows, from Lemma A.4.12 and the Riemann criterion, that \( \int_{\mathbb{R}^m} f (\cdot, y) \, dV_y \in \mathcal{R} (\mathbb{R}^n) \).

It remains to verify 1.27. First note

\[
\int_{\mathbb{R}^{n+m}} f (x) \, dV \in [L_{\mathcal{G}} (f), U_{\mathcal{G}} (f)].
\]

Next,
\[
L_{\mathcal{G}} (f) \leq \int_{\mathbb{R}^{n+m}} \psi \, dV = \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} \psi dV_y dV_x \leq \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} f (x, y) \, dV_y dV_x \leq \int_{\mathbb{R}^{n+m}} \phi \, dV \leq U_{\mathcal{G}} (f).
\]

Therefore,
\[
\left| \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} f (x, y) \, dV_y dV_x - \int_{\mathbb{R}^{n+m}} f (x) \, dV \right| \leq \varepsilon
\]
and since \( \varepsilon > 0 \) is arbitrary, this proves Fubini’s theorem\(^2\).

**Corollary A.5.5** Suppose \( E \) is a bounded contented set in \( \mathbb{R}^n \) 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 \( f \chi_P \in \mathcal{R} (\mathbb{R}^{n+1}) \) and

\[
\int_{\mathcal{P}} f \, dV = \int_{E} \int_{\psi (x)}^{\phi (x)} f (x, y) \, dy \, dV_x.
\]

**Proof:** Since \( f \) is continuous, there is no problem in writing \( f (x, \cdot) \chi_{[\psi (x), \phi (x)]} (\cdot) \in \mathcal{R} (\mathbb{R}^1) \). Also, \( f \chi_P \in \mathcal{R} (\mathbb{R}^{n+1}) \) because \( P \) is contented thanks to Corollary A.4.9. Therefore, by Fubini’s theorem

\[
\int_{\mathcal{P}} f \, dV = \int_{\mathbb{R}^n} \int_{\mathbb{R}} f \chi_P \, dy \, dV_x = \int_{E} \int_{\psi (x)}^{\phi (x)} f (x, y) \, dy \, dV_x
\]
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.
A.6 The Change Of Variables Formula

First recall Theorem A.6.1 on Page 607 which is listed here for convenience.

**Theorem A.6.1** Let \( h : U \rightarrow \mathbb{R}^n \) be a \( C^1 \) function with \( h(0) = 0, Dh(0)^{-1} \) exists. Then there exists an open set, \( V \subseteq U \) containing 0, flips, \( F_1, \ldots, F_{n-1} \), and primitive functions, \( G_n, G_{n-1}, \ldots, G_1 \) such that for \( x \in V \),

\[
h(x) = F_1 \circ \cdots \circ F_{n-1} \circ G_n \circ \cdots \circ G_1 (x).
\]

Also recall Theorem A.6.4 on Page 608.

**Theorem A.6.2** Let \( \phi : [a,b] \rightarrow [c,d] \) be one to one and suppose \( \phi' \) exists and is continuous on \( [a,b] \). Then if \( f \) is a continuous function defined on \( [a,b] \),

\[
\int_c^d f(s) \, ds = \int_a^b f(\phi(t)) |\phi'(t)| \, dt
\]

The following is a simple corollary to this theorem.

**Corollary A.6.3** Let \( \phi : [a,b] \rightarrow [c,d] \) be one to one and suppose \( \phi' \) exists and is continuous on \( [a,b] \). Then if \( f \) is a continuous function defined on \( [a,b] \),

\[
\int \chi_{[a,b]} (\phi^{-1} (x)) f(x) \, dx = \int \chi_{[a,b]} (t) f(\phi(t)) |\phi'(t)| \, dt
\]

**Lemma A.6.4** Let \( h : V \rightarrow \mathbb{R}^n \) be a \( C^1 \) function and suppose \( H \) is a compact subset of \( V \). Then there exists a constant, \( C \) independent of \( x \in H \) such that

\[
|Dh(x)| \leq C |v|.
\]

**Proof:** Consider the compact set, \( H \times \partial B(0,1) \subseteq \mathbb{R}^{2n} \). Let \( f : H \times \partial B(0,1) \rightarrow \mathbb{R} \) be given by \( f(x,v) = |Dh(x)v| \). Then let \( C \) denote the maximum value of \( f \). It follows that for \( v \in \mathbb{R}^n \),

\[
|Dh(x)| |v| \leq C
\]

and so the desired formula follows when you multiply both sides by \( |v| \).

**Definition A.6.5** Let \( A \) be an open set. Write \( C^k(A;\mathbb{R}^n) \) to denote a \( C^k \) function whose domain is \( A \) and whose range is \( \mathbb{R}^n \). Let \( U \) be an open set in \( \mathbb{R}^n \). Then \( h \in C^k(\overline{U};\mathbb{R}^n) \) if there exists an open set, \( V \supseteq U \) and a function, \( g \in C^1(V;\mathbb{R}^n) \) such that \( g = h \) on \( U \). \( f \in C^k(\overline{U}) \) means the same thing except that \( f \) has values in \( \mathbb{R} \).

**Theorem A.6.6** Let \( U \) be a bounded open set such that \( \partial U \) has zero content and let \( h \in C(\overline{U};\mathbb{R}^n) \) be one to one and \( Dh(x)^{-1} \) exists for all \( x \in U \). Then \( h(\partial U) = \partial(h(U)) \) and \( \partial(h(U)) \) has zero content.

**Proof:** Let \( x \in \partial U \) and let \( g = h \) where \( g \) is a \( C^1 \) function defined on an open set containing \( U \). By the inverse function theorem, \( g \) is locally one to one and an open mapping near \( x \). Thus \( g(x) = h(x) \) and is in an open set containing points of \( g(U) \) and points of \( g(U^c) \). These points of \( g(U^c) \) cannot equal any points of \( h(U) \) because \( g \) is one to one locally. Thus \( h(x) \in \partial(h(U)) \) and so \( h(\partial U) \subseteq \partial(h(U)) \). Now suppose \( y \in \partial(h(U)) \). By the inverse function theorem \( y \) cannot be in the open set \( h(U) \). Since \( y \in \partial(h(U)) \), every ball centered at \( y \) contains points of \( h(U) \) and so \( y \in \overline{h(U)} \setminus h(U) \). Thus there exists
a sequence, \( \{ x_n \} \subseteq U \) such that \( h(x_n) \to y \). But then, by the inverse function theorem, \( x_n \to h^{-1}(y) \) and so \( h^{-1}(y) \in \partial U \). Therefore, \( y \in h(\partial U) \) and this proves the two sets are equal. It remains to verify the claim about content.

First let \( H \) denote a compact set whose interior contains \( \overline{U} \) which is also in the interior of the domain of \( g \). Now since \( \partial U \) has content zero, it follows that for \( \varepsilon > 0 \) given, there exists a grid, \( G' \) such that if \( G' \) are those boxes of \( G \) which have nonempty intersection with \( \partial U \), then

\[
\sum_{Q \in G'} v(Q) < \varepsilon.
\]

and by refining the grid if necessary, no box of \( G \) has nonempty intersection with both \( U \) and \( H^c \). Refining this grid still more, you can also assume that for all boxes in \( G' \),

\[
l_i \leq 2
\]

where \( l_i \) is the length of the \( i \)th side. (Thus the boxes are not too far from being cubes.)

Let \( C \) be the constant of Lemma A.6.4 applied to \( g \) on \( H \).

Now consider one of these boxes, \( Q \in G' \). If \( x, y \in Q \), it follows from the chain rule that

\[
g(y) - g(x) = \int_0^1 Dg(x + t(y - x)) (y - x) \, dt
\]

By Lemma A.6.4 applied to \( H \)

\[
|g(y) - g(x)| \leq \int_0^1 |Dg(x + t(y - x))| |y - x| \, dt
\]

\[
\leq C \int_0^1 |x - y| \, dt \leq C \text{diam}(Q)
\]

\[
= C \left( \sum_{i=1}^n l_i^2 \right)^{1/2} \leq C \sqrt{n}L
\]

where \( L \) is the length of the longest side of \( Q \). Thus \( \text{diam}(g(Q)) \leq C \sqrt{n}L \) and so \( g(Q) \) is contained in a cube having sides equal to \( C \sqrt{n}L \) and volume equal to

\[
C^n n^{n/2} L^n \leq C^n n^{n/2} 2^n l_1 l_2 \cdots l_n = C^n n^{n/2} 2^n v(Q).
\]

Denoting by \( P_Q \) this cube, it follows

\[
h(\partial U) \subseteq \bigcup_{Q \in G'} v(P_Q)
\]

and

\[
\sum_{Q \in G'} v(P_Q) \leq C^n n^{n/2} 2^n \sum_{Q \in G'} v(Q) < \varepsilon C^n n^{n/2} 2^n.
\]

Since \( \varepsilon > 0 \) is arbitrary, this shows \( h(\partial U) \) has content zero as claimed.

**Theorem A.6.7** Suppose \( f \in C(\overline{U}) \) where \( U \) is a bounded open set with \( \partial U \) having content 0. Then \( f \mathcal{X}_U \in \mathcal{R}(\mathbb{R}^n) \).

**Proof:** Let \( H \) be a compact set whose interior contains \( \overline{U} \) which is also contained in the domain of \( g \) where \( g \) is a continuous functions whose restriction to \( U \) equals \( f \). Consider \( g \mathcal{X}_U \), a function whose set of discontinuities has content 0. Then \( g \mathcal{X}_U = f \mathcal{X}_U \in \mathcal{R}(\mathbb{R}^n) \) as claimed. This is by the big theorem which tells which functions are Riemann integrable.

The following lemma is obvious from the definition of the integral.
A.6. THE CHANGE OF VARIABLES FORMULA

Lemma A.6.8 Let \( U \) be a bounded open set and let \( f \mathcal{X}_U \in \mathcal{R} (\mathbb{R}^n) \). Then

\[
\int f (x + p) \mathcal{X}_{U - p} (x) \, dx = \int f (x) \mathcal{X}_U (x) \, dx.
\]

A few more lemmas are needed.

Lemma A.6.9 Let \( S \) be a nonempty subset of \( \mathbb{R}^n \). Define

\[ f (x) \equiv \text{dist} (x, S) \equiv \inf \{|x - y| : y \in S\}. \]

Then \( f \) is continuous.

Proof: Consider \(|f (x) - f (x_1)|\) and suppose without loss of generality that \( f (x_1) \geq f (x) \). Then choose \( y \in S \) such that \( f (x) + \varepsilon > |x - y| \). Then

\[
|f (x_1) - f (x)| = f (x_1) - f (x) \leq f (x_1) - |x - y| + \varepsilon \\
\leq |x_1 - y| - |x - y| + \varepsilon \\
\leq |x_1 - x| + |x - y| - |x - y| + \varepsilon \\
= |x - x_1| + \varepsilon.
\]

Since \( \varepsilon \) is arbitrary, it follows that \(|f (x_1) - f (x)| \leq |x - x_1|\) and this proves the lemma.

Theorem A.6.10 (Urysohn’s lemma for \( \mathbb{R}^n \)) Let \( H \) be a closed subset of an open set, \( U \). Then there exists a continuous function, \( g : \mathbb{R}^n \rightarrow [0, 1] \) such that \( g(x) = 1 \) for all \( x \in H \) and \( g(x) = 0 \) for all \( x \notin U \).

Proof: If \( x \notin C \), a closed set, then \( \text{dist} (x, C) > 0 \) because if not, there would exist a sequence of points of \( C \) converging to \( x \) and it would follow that \( x \in C \). Therefore, \( \text{dist} (x, H) + \text{dist} (x, U^C) > 0 \) for all \( x \in \mathbb{R}^n \). Now define a continuous function, \( g \) as

\[ g (x) \equiv \frac{\text{dist} (x, U^C)}{\text{dist} (x, H) + \text{dist} (x, U^C)}. \]

It is easy to see this verifies the conclusions of the theorem and this proves the theorem.

Definition A.6.11 Define \( \text{spt}(f) \) (support of \( f \)) to be the closure of the set \( \{x : f(x) \neq 0\} \). If \( V \) is an open set, \( C_c (V) \) will be the set of continuous functions \( f \), defined on \( \mathbb{R}^n \) having \( \text{spt}(f) \subseteq V \).

Definition A.6.12 If \( K \) is a compact subset of an open set, \( V \), then \( K \prec \phi \prec V \) if \( \phi \in C_c (V) \), \( \phi (K) = \{1\} \), \( \phi (\mathbb{R}^n) \subseteq [0, 1] \).

Also for \( \phi \in C_c (\mathbb{R}^n) \), \( K \prec \phi \) if

\[ \phi (\mathbb{R}^n) \subseteq [0, 1] \text{ and } \phi (K) = 1. \]

and \( \phi \prec V \) if

\[ \phi (\mathbb{R}^n) \subseteq [0, 1] \text{ and } \text{spt}(\phi) \subseteq V. \]
Theorem A.6.13 (Partition of unity) Let $K$ be a compact subset of $\mathbb{R}^n$ and suppose

$$K \subseteq V = \bigcup_{i=1}^n V_i, \ V_i \text{ open.}$$

Then there exist $\psi_i \prec V_i$ with

$$\sum_{i=1}^n \psi_i(x) = 1$$

for all $x \in K$.

Proof: Let $K_1 = K \setminus \bigcup_{i=2}^n V_i$. Thus $K_1$ is compact because it is the intersection of a closed set with a compact set and $K_1 \subseteq V_1$. Let $K_1 \subseteq W_1 \subseteq \overline{W}_1 \subseteq V_1$ with $\overline{W}_1$ compact. To obtain $W_1, W_2, \ldots, W_n$ covers $K$ and $\overline{W}_1 \subseteq V_1$. Let $K_2 = K \setminus (\bigcup_{i=3}^n V_i \cup W_1)$. Then $K_2$ is compact and $K_2 \subseteq V_2$. Let $K_2 \subseteq W_2 \subseteq \overline{W}_2 \subseteq V_2 \overline{W}_2$ compact. Continue this way finally obtaining $W_1, \ldots, W_n$, $K \subseteq W_1 \cup \cdots \cup W_n$, and $\overline{W}_i \subseteq V_i \overline{W}_i$ compact. Now let $\overline{W}_i \subseteq U_i \subseteq \overline{U}_i \subseteq V_i, \overline{U}_i$ compact.

By Theorem A.6.10, there exist functions, $\phi_i, \gamma$ such that $\overline{U}_i \prec \phi_i \prec V_i$, $\bigcup_{i=1}^n W_i \prec \gamma \prec \bigcup_{i=1}^n \overline{U}_i$. Define

$$\psi_i(x) = \begin{cases} \gamma(x) \phi_i(x) / \sum_{j=1}^n \phi_j(x) & \text{if } \sum_{j=1}^n \phi_j(x) \neq 0, \\ 0 & \text{if } \sum_{j=1}^n \phi_j(x) = 0. \end{cases}$$

If $x$ is such that $\sum_{j=1}^n \phi_j(x) = 0$, then $x \notin \bigcup_{i=1}^n \overline{U}_i$. Consequently $\gamma(y) = 0$ for all $y$ near $x$ and so $\psi_i(y) = 0$ for all $y$ near $x$. Hence $\psi_i$ is continuous at such $x$. If $\sum_{j=1}^n \phi_j(x) \neq 0$, this situation persists near $x$ and so $\psi_i$ is continuous at such points. Therefore $\psi_i$ is continuous. If $x \in K$, then $\gamma(x) = 1$ and so $\sum_{j=1}^n \psi_j(x) = 1$. Clearly $0 \leq \psi_i(x) \leq 1$ and $\text{spt}(\psi_j) \subseteq V_j$. This proves the theorem.

The next lemma contains the main ideas.

Lemma A.6.14 Let $U$ be a bounded open set with $\partial U$ having content 0. Also let $h \in C^1(\overline{U}; \mathbb{R}^n)$ be one to one on $U$ with $Dh(x)^{-1}$ exists for all $x \in U$. Let $f \in C(\overline{U})$ be nonnegative. Then

$$\int_{h(U)} (z) f(z) dV_n = \int_{U} f(h(x)) \left| \det Dh(x) \right| dV_n$$

Proof: Let $\varepsilon > 0$ be given. Then by Theorem A.6.7,

$$x \rightarrow \chi_U(x) f(h(x)) \left| \det Dh(x) \right|$$

is Riemann integrable. Therefore, there exists a grid, $\mathcal{G}$ such that, letting

$$g(x) = \chi_U(x) f(h(x)) \left| \det Dh(x) \right|,$$

$$\mathcal{L}_\mathcal{G}(g) + \varepsilon > U_\mathcal{G}(g).$$
A.6. THE CHANGE OF VARIABLES FORMULA

Let \( K \) denote the union of the boxes, \( Q \) of \( G \) for which \( m_Q (g) > 0 \). Thus \( K \) is a compact subset of \( U \) and it is only the terms from these boxes which contribute anything nonzero to the lower sum. By Theorem 4.2.4 on Page 660 which is stated above, it follows that for \( p \in K \), there exists an open set contained in \( U \) which contains \( p, O_p \) such that for \( x \in O_p - p \),

\[
\mathbf{h} (x + p) - \mathbf{h} (p) = \mathbf{F}_1 \circ \cdots \circ \mathbf{F}_{n-1} \circ \mathbf{G}_n \circ \cdots \circ \mathbf{G}_1 (x)
\]

where the \( \mathbf{G}_i \) are primitive functions and the \( \mathbf{F}_j \) are flips. Finitely many of these open sets, \( \{ O_j \}_{j=1}^m \) cover \( K \). Let the distinguished point for \( O_j \) be denoted by \( \mathbf{p}_j \). Now refine \( G \) if necessary such that the diameter of every cell of the new \( G \) which intersects \( U \) is smaller than a Lebesgue number for this open cover. Denote by \( G' \) those boxes of \( G \) whose union equals the set, \( K \). Thus every box of \( G' \) is contained in one of these \( O_j \). By Theorem A.6.13 on Page 661 there exists a partition of unity, \( \{ \psi_j \} \) on \( \mathbf{h} (K) \) such that \( \psi_j \prec \mathbf{h} (O_j) \). Then

\[
\mathcal{L}_{\mathcal{G}} (g) \leq \sum_{Q \in \mathcal{G}'} \int_{\partial Q} \mathcal{X}_Q (x) \psi (\mathbf{h} (x)) |\det \mathbf{h} (x)| \, dx
\]

\[
= \sum_{Q \in \mathcal{G}'} \sum_{j=1}^q \int_{\partial Q} \mathcal{X}_Q (x) \psi_j (\mathbf{h} (x)) |\det \mathbf{h} (x)| \, dx.
\]

(1.28)

Consider the term \( \int \mathcal{X}_Q (x) (\psi_j f) (\mathbf{h} (x)) \left( |\det \mathbf{h} (x)| \right) \, dx \). By Lemma A.6.5 and Fubini’s theorem this equals

\[
\int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} \mathcal{X}_{Q-p_j} (x) (\psi_j f) (\mathbf{h} (p) - \mathbf{F}_1 \circ \cdots \circ \mathbf{F}_{n-1} \circ \mathbf{G}_n \circ \cdots \circ \mathbf{G}_1 (x)) \cdot
\]

\[
|D\mathbf{F} (\mathbf{G}_n) \circ \cdots \circ \mathbf{G}_1 (x))| |D\mathbf{G}_n (\mathbf{G}_{n-1} \circ \cdots \circ \mathbf{G}_1 (x))| |D\mathbf{G}_{n-1} (\mathbf{G}_{n-2} \circ \cdots \circ \mathbf{G}_1 (x))| \cdots |D\mathbf{G}_2 (\mathbf{G}_1 (x))| \, dx \, dV_{n-1}.
\]

(1.29)

Here \( dV_{n-1} \) is with respect to the variables, \( x_2, \cdots, x_n \). Also \( F \) denotes \( \mathbf{F}_1 \circ \cdots \circ \mathbf{F}_{n-1} \). Now

\[
\mathbf{G}_1 (x) = (\alpha (x), x_2, \cdots, x_n)^T
\]

and is one to one. Therefore, fixing \( x_2, \cdots, x_n, x_1 \to \alpha (x) \) is one to one. Also, \( D\mathbf{G}_1 (x) = \frac{\partial }{\partial x_1} (x) \). Fixing \( x_2, \cdots, x_n \), change the variable,

\[
y_1 = \alpha (x_1, x_2, \cdots, x_n).
\]

Thus

\[
x = (x_1, x_2, \cdots, x_n)^T = \mathbf{G}_1^{-1} (y_1, x_2, \cdots, x_n) \equiv \mathbf{G}_1^{-1} (x')
\]

Then in (1.28) you can use Corollary A.6.3 to write (1.28) as

\[
\int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} \mathcal{X}_{Q-p_j} (G_1^{-1} (x')) (\psi_j f) (\mathbf{h} (p) - \mathbf{F}_1 \circ \cdots \circ \mathbf{F}_{n-1} \circ \mathbf{G}_n \circ \cdots \circ \mathbf{G}_1 (G_1^{-1} (x'))) \cdot
\]

\[
|D\mathbf{F} (\mathbf{G}_n \circ \cdots \circ \mathbf{G}_1 (G_1^{-1} (x')))\circ \cdots \circ \mathbf{G}_1 (G_1^{-1} (x'))) | \, dy_1 \, dV_{n-1}.
\]

(1.30)

which reduces to

\[
\int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} \mathcal{X}_{Q-p_j} (G_1^{-1} (x')) (\psi_j f) (\mathbf{h} (p) - \mathbf{F}_1 \circ \cdots \circ \mathbf{F}_{n-1} \circ \mathbf{G}_n \circ \cdots \circ \mathbf{G}_2 (x')) \cdot
\]

\[
|D\mathbf{F} (\mathbf{G}_n \circ \cdots \circ \mathbf{G}_2 (x'))\circ \cdots \circ \mathbf{G}_2 (x')) | \, dy_1 \, dV_{n-1}.
\]

(1.31)
Now use Fubini’s theorem again to make the inside integral taken with respect to \( x_2 \). Exactly the same process yields

\[
\int_{R^{n-1}} \int_R \mathcal{X}_{Q-p_j} \left( G_1^{-1} \circ G_2^{-1}(x'') \right) (\psi_j f) (h(p_i) + F_1 \circ \cdots \circ F_{n-1} \circ G_n \circ \cdots \circ G_3(x'')) \cdot |DF \left( G_n \circ \cdots \circ G_3(x'') \right)| \cdot |DG_n \left( G_{n-1} \circ \cdots \circ G_3(x'') \right)| \cdot |DG_{n-1} \left( G_{n-2} \circ \cdots \circ G_3(x'') \right)| \cdot \cdots dy_2 dV_{n-1}.
\]

(1.32)

Now \( F \) is just a composition of flips and so \( |DF \left( G_n \circ \cdots \circ G_3(x'') \right)| = 1 \) and so this term can be replaced with 1. Continuing this process, eventually yields an expression of the form

\[
\int_R \mathcal{X}_{Q-p_j} \left( G_1^{-1} \circ \cdots \circ G_{n-2}^{-1} \circ G_{n-1}^{-1} \circ G_n^{-1} \circ F^{-1}(y) \right) (\psi_j f) (h(p_i) + y) dV_n.
\]

(1.33)

Denoting by \( G^{-1} \) the expression, \( G_1^{-1} \circ \cdots \circ G_{n-2}^{-1} \circ G_{n-1}^{-1} \circ G_n^{-1}, \)

\[
\mathcal{X}_{Q-p_j} \left( G_1^{-1} \circ \cdots \circ G_{n-2}^{-1} \circ G_{n-1}^{-1} \circ G_n^{-1} \circ F^{-1}(y) \right) = 1
\]

exactly when \( G^{-1} \circ F^{-1}(y) \in Q - p_j \). Now recall that \( h(p_j + x) - h(p_j) = F \circ G(x) \) and so the above holds exactly when

\[
y = h(p_j + G^{-1} \circ F^{-1}(y)) - h(p_j) \in h(p_j + Q - p_j) - h(p_j)
\]

\[
= h(Q) - h(p_j).
\]

Thus \( \mathcal{U} \) reduces to

\[
\int_R \mathcal{X}_{h(Q) - h(p_j)} (y) (\psi_j f) (h(p_i) + y) dV_n = \int_R \mathcal{X}_{h(Q)} (z) (\psi_j f) (z) dV_n.
\]

It follows from \( \mathcal{U} \)

\[
\mathcal{U}_G (g) - \varepsilon \leq \mathcal{L}_G (g) \leq \sum_{Q \in \mathcal{P}'} \int \mathcal{X}_Q (x) f(h(x)) |\det Dh(x)| dx
\]

\[
= \sum_{Q \in \mathcal{P}'} \sum_{j=1}^{q} \int \mathcal{X}_Q (x) (\psi_j f) (h(x)) |\det Dh(x)| dx
\]

\[
= \sum_{Q \in \mathcal{P}'} \sum_{j=1}^{q} \int_R \mathcal{X}_{h(Q)} (z) (\psi_j f) (z) dV_n
\]

\[
= \sum_{Q \in \mathcal{P}'} \int_R \mathcal{X}_{h(Q)} (z) f(z) dV_n \leq \int \mathcal{X}_{h(U)} (z) f(z) dV_n
\]

which implies the inequality,

\[
\int \mathcal{X}_U (x) f(h(x)) |\det Dh(x)| dV_n \leq \int \mathcal{X}_{h(U)} (z) f(z) dV_n
\]

But now you can use the same information just derived to obtain equality. \( x = h^{-1}(z) \) and so from what was just done,

\[
\int \mathcal{X}_U (x) f(h(x)) |\det Dh(x)| dV_n
\]
\[
\begin{align*}
\int \mathcal{X}_{h^{-1}(U)}(x) f(h(x)) |\det D h(x)| dV_n \\
&\geq \int \mathcal{X}_{h(U)}(z) f(z) |\det D h^{-1}(z)| |\det D h^{-1}(z)| dV_n \\
&= \int \mathcal{X}_{h(U)}(z) f(z) dV_n
\end{align*}
\]
from the chain rule. In fact,
\[
I = \det D h^{-1}(z) D h^{-1}(z)
\]
and so
\[
1 = |\det D h^{-1}(z)| |\det D h^{-1}(z)|.
\]
This proves the lemma.

The change of variables theorem follows.

**Theorem A.6.15** Let \( U \) be a bounded open set with \( \partial U \) having content 0. Also let \( h \in C^1(U; \mathbb{R}^n) \) be one to one on \( U \) with \( Dh(x)^{-1} \) exists for all \( x \in U \). Let \( f \in C(U) \). Then
\[
\int \mathcal{X}_{h(U)}(z) f(z) dz = \int \mathcal{X}_U(x) f(h(x)) |\det D h(x)| dx
\]

**Proof:** You note that the formula holds for \( f^+ = \frac{1+f}{2} \) and \( f^- = \frac{1-f}{2} \). Now \( f = f^+ - f^- \) and so
\[
\begin{align*}
\int \mathcal{X}_{h(U)}(z) f(z) dz &= \int \mathcal{X}_{h(U)}(z) f^+(z) dz - \int \mathcal{X}_{h(U)}(z) f^-(z) dz \\
&= \int \mathcal{X}_U(x) f^+(h(x)) |\det D h(x)| dx - \int \mathcal{X}_U(x) f^-(h(x)) |\det D h(x)| dx \\
&= \int \mathcal{X}_U(x) f(h(x)) |\det D h(x)| dx.
\end{align*}
\]

### A.7 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. However, most of the difficulties are artifacts. They shouldn’t even be considered! 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 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. An introduction to it is in [21]. Another very good source is [14]. 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 [13], [15], [22], and [16]. There is also a still more general integral called the generalized Riemann integral. A recent book on this subject is [11]. It is far easier to define than the Lebesgue integral but the convergence theorems are much harder to prove. An introduction is also in [13].
Bibliography


Index

$C^1$, $C^k$, $C^\infty$, $\Delta$, $\nabla^2$

adjugate, 107, 166
agony, pain and suffering, 414
algebraic multiplicity, 234
angle between planes, 77
angle between vectors, 45
angular velocity, 49
angular velocity vector, 302
arc length, 263, 297, 473
area of a parallelogram, 48
augmented matrix, 87
back substitution, 86
balance of momentum, 533
barallepiped
   volume, 51
basic variables, 19
basis, 186
binormal, 291
block matrix, 136
bounded, 333
box product, 51
Cartesian coordinates, 15
Catenary, 349
Cauchy Schwarz inequality, 40
Cauchy sequence, 336
Cauchy sequence, 437
Cauchy stress, 437
Cayley Hamilton theorem, 229
center of mass, 56, 462, 464
center of mass of a plate, 117
center of mass of a surface, 507
centroid, 62
chain rule, 393
change of variables formula, 449
characteristic equation, 228
characteristic polynomial, 183
characteristic value, 228
Circular helix, 291
classical adjoint, 167
closed set, 312
cofactor, 162, 180
cofactor matrix, 180
column space, 406
complement, 312
complex eigenvalues, 248
component, 56, 152
component of a force, 43
components of a matrix, 180
conformal, 107
conservation of linear momentum, 478
conservation of mass, 532
conservative, 538, 107
constitutive laws, 538
contented set, 601
continuity
   limit of a sequence, 608
continuous function, 417
continuous functions
   properties, 316
converge, 606
Coordinates, 17
Cramer’s rule, 171
critical point, 363
cross product, 48
Curl, 485
Curvature, 283, 290
defective, 235
defective eigenvalue, 235
definition of a helix, 291
deformation gradient, 533
density and mass, 416
density with respect to area, 507
dependence relation, 130, 132
circular helix, 491
classical adjoint, 167
closed set, 312
cofactor, 162, 180
cofactor matrix, 180
column space, 406
complement, 312
complex eigenvalues, 248
component, 56, 152
component of a force, 43
components of a matrix, 180
conformal, 107
conservation of linear momentum, 478
conservation of mass, 532
conservative, 538, 107
constitutive laws, 538
contented set, 601
continuity
   limit of a sequence, 608
continuous function, 417
continuous functions
   properties, 316
converge, 606
Coordinates, 17
Cramer’s rule, 171, 182
critical point, 363
cross product, 48
Curl, 485
Curvature, 283, 290
defective, 235
defective eigenvalue, 235
definition of a helix, 291
definition of a point, 363
deformation gradient, 533
density and mass, 416
density with respect to area, 507
dependence relation, 130, 132
derivative of a function, 261
determinant, 175
alternating property, 177
cofactor expansion, 180
expansion along row (column), 180
matrix inverse formula, 181
product, 179
transpose, 177
diagonalizable, 238
diameter, 333
difference quotient, 261
differentiable matrix, 299
differentiation rules, 266
digraph, 107
dimension of vector space, 199
directed graph, 107
directed line segment, 28
direction vector, 26
directional derivative, 323
distance formula, 22
divergence, 485
 divergence theorem, 523
Dolittle's method, 153
domain, 260
 dominant eigenvalue, 231
donut, 505
dot product, 37
echelon form, 88
eigenspace, 228
eigenvalue, 228
eigenvalues, 183
eigenvector, 228
Einstein summation convention, 59
elementary matrices, 122
 elementary matrix
 inverse, 122
 properties, 122
 entries of a matrix, 111
equality of mixed partial derivatives, 329
Eulerian coordinates, 533
Fibonacci sequence, 335
force, 31
force field, 178
free variables, 49
Frenet Serret formulas, 291
fundamental theorem line integrals, 413
Gauss Elimination, 96
Gauss elimination, 97
Gauss Jordan method for inverses, 116
Gauss-Seidel method, 573
Gauss’s theorem, 240
general solution, 224
geometric multiplicity, 228
Gerschgorin’s theorem, 248
gradient, 326
Green’s theorem, 493
 grid, 111, 113, 124, 126
 grids, 108
Heine Borel, 294
Hessian matrix, 365
homogeneous equations, 142
implicit function theorem, 399
impulse, 278
inconsistent, 97, 94
increment of area, 449
increment of volume
 increment of area, 449
inner product, 37
intercepts, 77
intercepts of a surface, 610
interior point, 448
inverse
 left inverse, 184
 right inverse, 184
inverses and determinants, 112, 113
invertible, 112
iterated integrals, 415
Jacobian, 449
 Jacobian determinant, 449
Jocobi method, 569
Jordan content, 598
 Jordan set, 601
 joule, 46
ker, 222
kernel, 222
kilogram, 56
kinetic energy, 277
Kronecker delta, 59
 Lagrange multipliers, 373
Lagrange remainder, 395
Lagrangian coordinates, 533
Laplace expansion, 162
Laplace equation, 180
leading entry, 88
leading variables, 49
Lebesgue number, 601
Lebesgue’s theorem, 601
Lebesgue’s theorem, 108
INDEX

length of smooth curve, 264
limit of a function, 260, 319
limits and continuity, 261
line integral, 480
linear combination, 129, 130, 178, 194
linear momentum, 278
linear relation, 129
linear relationship, 129
linear space, 190
linear transformation, 212, 380
linearly independent, 196
lizards
surface area, 424, 502
local extremum, 362
local maximum, 362
local minimum, 362
lower sum, 427, 594
main diagonal, 166
mass balance, 532
material coordinates, 533
matrix
inverse, 112
left inverse, 182
lower triangular, 163, 183
polynomial, 184
right inverse, 182
self adjoint, 254
symmetric, 254
upper triangular, 163, 183
migration matrix, 245
minor, 164, 180
mixed partial derivatives, 328
moment of a force, 54
moment of inertia, 463
motion, 533
moving coordinate system, 299
multi-index, 316
nested interval lemma, 614
Newton
second law, 273
Newton's laws, 273
nondefective eigenvalue, 235
normal vector to plane, 71
null space, 121, 192, 222
nullity, 205
one to one, 214
onto, 214
open set, 312
orientable, 554
orientation, 178
oriented curve, 173
origin, 17
orthogonal, 39
osculating plane, 283, 290
parallelepiped, 71
parameter, 28, 260
parametric equation, 28
parametrization, 260, 263, 297, 473
partial derivative, 325
partition of unity, 316
permutation, 177
permutation matrices, 122
permutation symbol, 59
perpendicular, 39
Piola Kirchhoff stress, 537
pivot, 89
pivot column, 89
pivot position, 89
plane containing three points, 63
planes, 71
polynomial
matrix coefficients, 188
polynomials in n variables, 316
position vector, 20, 31
power method, 579
precession of a top, 459
pretentious jargon
hyper plane, 79, 84
oid, 308
principal normal, 283, 290
product rule
cross product, 266
dot product, 266
matrices, 156
projection of a vector, 44
QR factorization, 156
quadric surfaces, 308
radius of curvature, 283, 290
rank, 142, 203
rank of a matrix, 208
rank theorem, 143
raw eggs, 463
recurrence relation, 335
recursively defined sequence, 335
refinement of a grid, 411, 426
refinement of grids, 593
resultant, 33
Riemann criterion, 595
Riemann integral, 411, 426
right handed system, 47
rot, 153, 251
row equivalent, 158
row operations, 122, 131, 164
row reduced echelon form, 133
row space, 203
saddle point, 365
scalar field, 485, 521
scalar multiplication, 18
scalar potential, 482, 558
scalar product, 37
scalars, 18, 99
scaling factor, 580
second derivative test, 397
separable differential equations, 346
sequences, 335
sequential compactness, 294, 337
sequentially compact, 337
sgn, 173
uniqueness, 174
shifted inverse power method, 585
sign of a permutation, 174
simultaneous corrections, 585
singular point, 363
skew lines, 83, 144
skew symmetric, 111
smooth curve, 263, 297, 473
smooth surface, 499
solution set, 85
solution space, 222
spacial coordinates, 533
span, 178, 194
spectrum, 228
speed, 34
spherical coordinates, 390
standard matrix, 380
standard position, 32
Stoke’s theorem, 551
subspace, 191
substituting matrix into polynomial identity, 184
support of a function, 613
symmetric, 111
symmetric form of a line, 55

trace of a surface, 510
traces, 21
triangle inequality, 24, 41
triple vector product, 76
unit tangent vector, 488, 511
unit vector, 26
upper sum, 427, 594

vector, 19
vector field, 476, 478, 485, 521
vector fields, 174
vector potential, 488
vector space, 190
vector valued function
continuity, 410
derivative, 410
integral, 410
limit theorems, 410
vectors, 19
velocity, 34
volume element, 449
volume increment, 449
work, 479