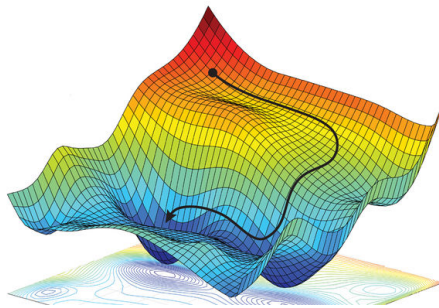


Math Refresher for Machine Learning



Paul F. Roysdon, Ph.D.

2019

MATH REFRESHER FOR MACHINE LEARNING

First Edition

Copyright © 2019 by Paul F. Roysdon, Ph.D. All rights reserved. Printed in the United States of America. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a data base or retrieval system, without the prior written permission of the publisher.

ISBN 9781093669077

The text for this book was formatted in *LATEX* and the mathematics was formatted in *AMS-LATEX* (Donald Knuth's *TEXtext* formatting system) and converted from device-independent to postscript format using *DVIPS*.

*To my Applied Mathematics & Data Science colleagues, thank
you.*

Contents

Preface	viii
Pre-Refresher Exercises	x
1 Introduction	1
1.1 Statistics	1
1.2 Calculus	2
1.3 Linear Algebra	3
1.4 Probability Theory	3
2 Linear Algebra	5
2.1 Working with Vectors	5
2.1.1 Vector	5
2.1.2 Vector Addition and Subtraction	5
2.1.3 Scalar Multiplication	5
2.1.4 Vector Inner Product	6
2.1.5 Vector Norm	6
2.2 Linear Independence	7
2.2.1 Linear combinations	7
2.2.2 Linear independence	7
2.3 Basics of Matrix Algebra	8
2.3.1 Matrix	8
2.3.2 Matrix Addition	8
2.3.3 Scalar Multiplication	9
2.3.4 Matrix Multiplication	9
2.3.5 Laws of Matrix Algebra	11
2.3.6 Transpose	11
2.3.7 Properties of the transpose	12
2.4 Systems of Linear Equations	13
2.4.1 Linear Equation	13
2.5 Systems of Equations as Matrices	14
2.5.1 Coefficient matrix	14

2.5.2	Augmented Matrix	15
2.6	Solutions to Augmented Matrices & Systems of Equations	15
2.6.1	Row Echelon Form	15
2.6.2	Reduced Row Echelon Form	16
2.6.3	Gaussian and Gauss-Jordan elimination	16
2.7	Rank & Number of Solutions	17
2.8	The Inverse of a Matrix	18
2.8.1	Properties of the Inverse	19
2.8.2	Procedure to Find the Inverse	19
2.9	Linear Systems and Inverses	21
2.10	Determinants	21
2.11	Matrix Inverse using the Determinant	23
3	Functions and Operations	25
3.1	Summation and Product Operators	25
3.1.1	Summation	25
3.1.2	Product	26
3.1.3	Factorials!	26
3.1.4	Modulo	26
3.2	Introduction to Functions	27
3.2.1	Functions	27
3.2.2	Dimensionality	27
3.2.3	Definitions	29
3.3	log and exponent	29
3.3.1	Relationship of logarithmic and exponential func- tions	29
3.3.2	Common Bases	29
3.3.3	Properties of exponential functions	30
3.3.4	Properties of logarithmic functions of any base . .	30
3.3.5	Change of Base Formula	30
3.3.6	Log, Product, and Sum Operators	31
3.4	Graphing Functions	32
3.5	Solving for Variables and Finding Roots	32
3.5.1	Procedure	33
3.5.2	Quadratic Formula	33
3.6	Sets	33
3.6.1	Interior Point	33
3.6.2	Boundary Point	34
3.6.3	Open	34
3.6.4	Closed	34
3.6.5	Complement	34
3.6.6	Empty	35

4	Limits	37
4.1	The Central Limit Theorem	37
4.2	The Law of Large Numbers	38
4.3	Sequences	38
4.4	The Limit of a Sequence	40
4.5	Limits of a Function	41
4.5.1	Properties of Limits	42
4.6	Continuity	44
4.6.1	Properties of Continuous Functions	44
5	Calculus	47
5.1	The Mean is a Type of Integral	47
5.2	Derivatives	48
5.2.1	Properties of derivatives	49
5.3	Higher-Order Derivatives (Derivatives of Derivatives) . . .	50
5.4	Composite Functions and the Chain Rule	50
5.5	Derivatives of natural logs and the exponent	52
5.5.1	Derivatives of natural exponential function	52
5.5.2	Derivatives of log	53
5.5.3	Pseudo-Math Proof	54
5.6	Partial Derivatives	55
5.7	Taylor Series Approximation	56
5.8	The Indefinite Integration	56
5.8.1	Common Rules of Integration	58
5.9	The Definite Integral: The Area under the Curve	59
5.9.1	Common Rules for Definite Integrals	61
5.10	Integration by Substitution	62
5.11	Integration by Parts	63
6	Optimization	65
6.1	Maxima and Minima	65
6.2	Concavity of a Function	67
6.2.1	Quadratic Forms	68
6.2.2	Definiteness of Quadratic Forms	69
6.3	FOC and SOC	69
6.3.1	First Order Conditions	70
6.3.2	Second Order Conditions	71
6.3.3	Definiteness and Concavity	72
6.4	Global Maxima and Minima	73
6.5	Constrained Optimization	75
6.5.1	Equality Constraints	76
6.6	Inequality Constraints	78
6.7	Kuhn-Tucker Conditions	81
6.8	Applications of Quadratic Forms	84

7	Probability Theory	85
7.1	Counting rules	85
7.1.1	Fundamental Theorem of Counting	85
7.1.2	Sampling Table	85
7.2	Sets	86
7.2.1	Set	86
7.2.2	Sample Space (S)	86
7.2.3	Event	87
7.2.4	Empty Set	87
7.2.5	Set operations	87
7.2.6	Properties of set operations	87
7.3	Probability	88
7.3.1	Probability Definitions: Formal and Informal . . .	88
7.3.2	Probability Distribution Function	88
7.3.3	Probability Operations	89
7.4	Conditional Probability and Bayes Law	90
7.4.1	Conditional Probability	90
7.4.2	Bayes Rule	92
7.4.3	Prior and Posterior Probabilities	92
7.5	Independence	92
7.5.1	Pairwise Independence	93
7.5.2	Conditional Independence	93
7.6	Random Variables	94
7.6.1	Randomness	94
7.7	Distributions	94
7.7.1	Discrete Random Variables	95
7.7.2	Continuous Random Variables	97
7.8	Joint Distributions	98
7.8.1	Marginal Probability Distribution	98
7.8.2	Conditional Probability Distribution	99
7.9	Expectation	99
7.9.1	Expected Value of a Discrete Random Variable . .	100
7.9.2	Expected Value of a Continuous Random Variable	100
7.9.3	Expected Value of a Function	100
7.9.4	Properties of Expected Values	101
7.9.5	Conditional Expectation	101
7.10	Variance and Covariance	102
7.10.1	Properties of Variance and Covariance	103
7.11	Special Distributions	104
7.11.1	Common Discrete distributions	104
7.11.2	Common Continuous Distributions	105
7.12	Summarizing Observed Events (Data)	107
7.12.1	Sample mean	107
7.12.2	Dispersion	107

7.12.3	Sample variance	107
7.12.4	Standard deviation	108
7.12.5	Covariance and Correlation	108
7.13	Asymptotic Theory	108
7.13.1	CLT and LLN	109
7.13.2	Big O-Notation	110
A	Conventions and Symbols	111
A.1	Notation	111
A.2	Acronyms	111
A.3	Greek letters	111
B	Solutions to Exercises	115
B.1	Solutions to Warm-up Questions	115
B.2	Solutions to Linear Algebra Exercises	119
B.3	Solutions to Functions and Operations Exercises	122
B.4	Solutions to Limits Exercises	123
B.5	Solutions to Calculus Exercises	124
B.6	Solutions to Optimization Exercises	128
B.7	Solutions to Probability Theory Exercises	130
	Index	135

Preface

Purpose. To provide a solid mathematical foundation for future work in Data Science (DS) topics, e.g. Machine Learning (ML), Regression, Data Analysis, etc.. This text is the ensemble of teaching notes provided by my colleagues at Harvard, Stanford, MIT, and CalTech, as well as my own notes gathered from years of mentoring and tutoring young aspiring engineers and data scientists. Unfortunately, some professionals enter the workplace without the requisite knowledge in math and programming skills to be contributing members of an engineering or data science team. Compounding this issue, there exists a large body of literature on this subject, much of which is confusing or inaccurate, with inconsistent nomenclature. This text is a review of high school and first-year college math, with applications in Matlab and Python, so that anyone can read this text and gain the skills necessary to:

- Perform research of new and exciting literature in ML.
- Apply their Matlab or Python skills to an application in their field of work.

Pre-requisites & Audience. None. This text and the accompanying lectures are for students interested in ML and DS. Most professionals have seen this material in high school or college. A reasonably prepared undergraduate student should be able to understand and apply all contents of this book. If some of the material is new, don't worry, this text will walk you through the necessary steps.

Structure & Requirements. In a classroom environment, this text can be covered in a two hour period over a course of eight days. However, this still requires a commitment from the student to read and complete each assignment. The lectures will focus on major topics and highlights, while the reading will cover all the material necessary to understand the ML concepts.

Computing & Associated Software. This text uses Matlab (or Octave) and Python, numerical computing packages with many free add-ons for linear algebra and machine learning, e.g. NumPy and PyTorch. We will use these tools extensively in our examples, so it is suggested that the student installs and familiarize themselves with the tools. You will not be required to be proficient in Matlab or Python, rather it is expected that you will learn tips and tricks throughout this text.

It is good to be able to solve small problems by hand, but in practice they are large, requiring a computer for their solution. Therefore, to fully appreciate the subject, one needs to solve large (practical) problems on

a computer. An important feature of this book is that it comes with software implementing the major algorithms described herein. At the time of writing, software for the following five algorithms is available:

- Least-Squares & Weighted Least Squares solvers.
- A basic Neural Network classifier.
- Linear Program (LP) & Integer Linear Program (ILP) solvers.
- A Monte Carlo tool.
- A Random Forest tool.

The programs that implement these algorithms are written in both Matlab (or Octave) and Python and can be easily run on most hardware platforms. Students/instructors are encouraged to install and run these programs on their local hardware. Great pains have been taken to make the source code for these programs readable. In particular, the names of the variables in the programs are consistent with the notation of this book.

The software can be downloaded from the following web site:
<https://github.com/AidedNav/ML>

Features. Here are some features that distinguish this book from others:

- The book gives a balanced treatment to both the traditional and newer methods. The notation and analysis is developed to be consistent across the methods.
- From the beginning and consistently throughout the book, example problems are formulated and solved to develop an understanding of the equations and their application to ML. By highlighting this throughout, it is hoped that the reader will more fully understand and appreciate theory behind ML.
- There is an extensive treatment of modern methods, including numerically fast solutions systems of equations.
- In addition to the traditional applications, which come mostly from business and economics, the book features other important applications.

Dependencies. In general, the later chapters depend on the notation and equations established in the earlier chapters.

P. F. Roysdon, Ph.D.
March 25, 2019

Preliminary Exercises

Before our first meeting, please try solving these questions. They are a sample of the very beginning of each math section. The questions are review of what you have seen in high-school and college. We have provided links to the parts of the book you can read if the concepts are new to you.

The goal of this “preliminary” assignment is not to intimidate you but to set common expectations so you can make the most out of the actual *Math Refresher for DS*. Even if you do not understand some or all of these questions after skimming through the linked sections, your effort will pay off and you will be better prepared for the math refresher.

Linear Algebra

Vectors

Define the vectors $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$, $\mathbf{v} = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}$, and the scalar $c = 2$.

Calculate the following:

1. $\mathbf{u} + \mathbf{v}$
2. $c\mathbf{v}$
3. $\mathbf{u} \cdot \mathbf{v}$

If you are having trouble with these problems, please review Section 2.1 “Working with Vectors” in Chapter 2.

Are the following sets of vectors linearly independent?

1. $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $\mathbf{v} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$
2. $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix}$, $\mathbf{v} = \begin{pmatrix} 3 \\ 7 \\ 9 \end{pmatrix}$
3. $\mathbf{a} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$, $\mathbf{b} = \begin{pmatrix} 3 \\ -4 \\ -2 \end{pmatrix}$, $\mathbf{c} = \begin{pmatrix} 5 \\ -10 \\ -8 \end{pmatrix}$ (this requires some guess-work)

If you are having trouble with these problems, please review Section 2.2.

Matrices

$$\mathbf{A} = \begin{pmatrix} 7 & 5 & 1 \\ 11 & 9 & 3 \\ 2 & 14 & 21 \\ 4 & 1 & 5 \end{pmatrix}$$

What is the dimensionality of matrix \mathbf{A} ?

What is the element a_{23} of \mathbf{A} ?

Given that

$$\mathbf{B} = \begin{pmatrix} 1 & 2 & 8 \\ 3 & 9 & 11 \\ 4 & 7 & 5 \\ 5 & 1 & 9 \end{pmatrix}$$

What is $\mathbf{A} + \mathbf{B}$?

Given that

$$\mathbf{C} = \begin{pmatrix} 1 & 2 & 8 \\ 3 & 9 & 11 \\ 4 & 7 & 5 \end{pmatrix}$$

What is $\mathbf{A} + \mathbf{C}$?

Given that

$$c = 2$$

What is $c\mathbf{A}$?

If you are having trouble with these problems, please review Section 2.3.

Operations

Summation

Simplify the following

1. $\sum_{i=1}^3 i$

2. $\sum_{k=1}^3 (3k + 2)$

3. $\sum_{i=1}^4 (3k + i + 2)$

Products

1. $\prod_{i=1}^3 i$
2. $\prod_{k=1}^3 (3k + 2)$

To review this material, please see Section 3.1.

Logs and exponents

Simplify the following

1. 4^2
2. $4^2 2^3$
3. $\log_{10} 100$
4. $\log_2 4$
5. $\log e$, where \log is the natural log (also written as \ln) – a log with base e , and e is Euler's constant
6. $e^a e^b e^c$, where a, b, c are each constants
7. $\log 0$
8. e^0
9. e^1
10. $\log e^2$

To review this material, please see Section 3.3.

Limits

Find the limit of the following.

1. $\lim_{x \rightarrow 2} (x - 1)$
2. $\lim_{x \rightarrow 2} \frac{(x-2)(x-1)}{(x-2)}$
3. $\lim_{x \rightarrow 2} \frac{x^2 - 3x + 2}{x - 2}$

To review this material please see Section 4.5.

Calculus

For each of the following functions $f(x)$, find the derivative $f'(x)$ or $\frac{d}{dx}f(x)$

1. $f(x) = c$

2. $f(x) = x$
3. $f(x) = x^2$
4. $f(x) = x^3$
5. $f(x) = 3x^2 + 2x^{1/3}$
6. $f(x) = (x^3)(2x^4)$

For a review, please see Section 5.2 - 5.3.

Optimization

For each of the following functions $f(x)$, does a maximum and minimum exist in the domain $x \in \mathbf{R}$? If so, for what are those values and for which values of x ?

1. $f(x) = x$
2. $f(x) = x^2$
3. $f(x) = -(x - 2)^2$

If you are stuck, please try sketching out a picture of each of the functions.

Probability

1. If there are 12 cards, numbered 1 to 12, and 4 cards are chosen, how many distinct possible choices are there? (unordered, without replacement)
2. Let $A = \{1, 3, 5, 7, 8\}$ and $B = \{2, 4, 7, 8, 12, 13\}$. What is $A \cup B$? What is $A \cap B$? If A is a subset of the Sample Space $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, what is the complement A^C ?
3. If we roll two fair dice, what is the probability that their sum would be 11?
4. If we roll two fair dice, what is the probability that their sum would be 12?

For a review, please see Sections 7.2 - 7.3.

Chapter 1

Introduction

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. There are four primary fields of mathematics commonly associated with data science:

- Statistics is used to extract useful information from data.
- Calculus tells us how to learn and optimize our model.
- Linear Algebra makes running these algorithms feasible on large datasets.
- Probability Theory helps us understand the likelihood of an event occurring.

1.1 Statistics

Statistics is a set of *techniques* that extract useful information from data, whereas statistical inference is the *process* of making a prediction about a larger population of data based on a smaller sample. Consider a simple 2-dimensional dataset with a normal (Gaussian) distribution, see Fig. 1.1

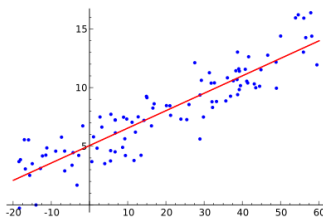


Figure 1.1: Random data line fit.

To fit the dataset with a line, we use a statistical inference technique called linear regression. This method allows us to summarize and study the relationship between two variables. One variable x is the independent variable, while y is the dependent variable. In this example we use the linear equation $y = ax + b$ for our linear regression, where y is the prediction, x is the input, b is the point where the line intersects the y -axis, and a is the slope of the line.

If a and b were known, then the line fit is simple. However, if the dataset is given and a and b are unknown, how could we find the “correct” line fit that best represents the data? The naïve way would be to try many different variables until we found a fit. Fortunately there is a way to find the optimal values that fit our dataset. To do this we need something called an error function. This error function will quantify the error between our values y , and our prediction \hat{y} .

There are many types of statistical error functions, but we will use a simple one called *least squares*. For n values, define the least squares error as

$$\delta y = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Numerically, first we make a prediction of what the values should be, then compute an error between the prediction and the data. Next we square each of those differences and sum them. This is the total error value. Next we make an adjustment, and make a new prediction. Picking the next adjustment, or “error” value, requires calculus.

1.2 Calculus

Calculus is the study of *change*. To solve for the next “error” value, we will use an optimization technique called *gradient descent*, that finds the minimum value, or minimum *cost* $\mathcal{J}(x)$, iteratively. At each iteration we compute the partial derivative $f'(x)$, or gradient, of our regression equation $f(x)$. The result is a *direction* of maximum change, *toward* our minimum. This process is repeated until a minimum is found. Consider, for example, Fig. 1.2. Calculus helps us find the direction of change most important to minimize our error and optimize our variables, in this case x , such that our prediction is most optimal, resulting in the smallest error. Practically, this method could continue for an infinite number of iterations, so some stopping criteria is often used when the error is below some pre-defined threshold.

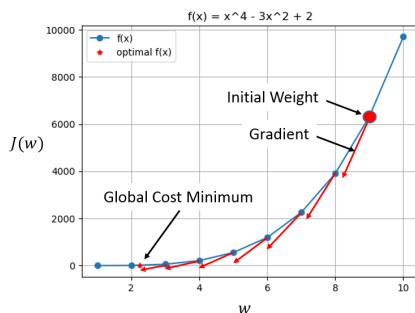


Figure 1.2: Gradient descent for a simple polynomial.

1.3 Linear Algebra

As we encounter more complex problems with more variables, the problem is a multi-variate problem. For example: instead of the single variable x in Fig. 1.2, a multi-variate polynomial could be $f(\mathbf{x}) = ax_1 + bx_2 + c$, where the variable is now a vector of values $\mathbf{x} = [x_1, x_2]$. Linear algebra is the field of *multi-variate spaces* and the linear *transformation* between them. Linear algebra provides a set of operations that we can perform on groups of numbers, known as matrices, as our training set becomes $m \times n$. For m samples that have n features, instead of a single variable with a weight, each feature has a weight.

1.4 Probability Theory

Probability is a *measure* of the likelihood of an event occurring. We can use a probabilistic technique called logistic regression to help us fit data into categories or classes. Instead of predicting a value, we are predicting the probability of an occurrence. Since the probability of an event is defined as values between 0 and 1, we cannot use an infinitely stretching line. We need a threshold function that passes some point x if the probability of an event is likely to occur. In machine learning, we often use an “s”-shaped curve like the Sigmoid function. Once we optimize our function using the methods above, we use probability theory to compute a probabilistic class value.

Chapter 2

Linear Algebra

2.1 Working with Vectors

2.1.1 Vector

Definition 2.1 (Vector). *A vector in n -space is an ordered list of n numbers.* ♠

These numbers can be represented as either a row vector or a column vector:

$$\mathbf{v} = (v_1 \quad v_2 \quad \dots \quad v_n), \quad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

We can also think of a vector as defining a point in n -dimensional space, usually \mathbb{R}^n ; each element of the vector defines the coordinate of the point in a particular direction.

2.1.2 Vector Addition and Subtraction

If two vectors, \mathbf{u} and \mathbf{v} , have the same length (i.e. have the same number of elements), they can be added (subtracted) together:

$$\mathbf{u} + \mathbf{v} = (u_1 + v_1 \quad u_2 + v_2 \quad \dots \quad u_k + v_n)$$

$$\mathbf{u} - \mathbf{v} = (u_1 - v_1 \quad u_2 - v_2 \quad \dots \quad u_k - v_n)$$

2.1.3 Scalar Multiplication

The product of a scalar c (i.e. a constant) and vector \mathbf{v} is:

$$c\mathbf{v} = (cv_1 \quad cv_2 \quad \dots \quad cv_n)$$

2.1.4 Vector Inner Product

The inner product (also called the dot product or scalar product) of two vectors \mathbf{u} and \mathbf{v} is again defined *iff* they have the same number of elements:

$$\mathbf{u} \cdot \mathbf{v} = u_1v_1 + u_2v_2 + \cdots + u_nv_n = \sum_{i=1}^n u_iv_i$$

If

$$\mathbf{u} \cdot \mathbf{v} = 0,$$

the two vectors are orthogonal (perpendicular).

2.1.5 Vector Norm

The norm of a vector is a measure of its length. There are many different ways to calculate the norm, but the most common is the Euclidean norm (which corresponds to our usual conception of distance in three-dimensional space):

$$\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{v_1v_1 + v_2v_2 + \cdots + v_nv_n}$$

Example 2.1 (Vector Algebra). Let $\mathbf{a} = \begin{pmatrix} 2 & 1 & 2 \end{pmatrix}$, $\mathbf{b} = \begin{pmatrix} 3 & 4 & 5 \end{pmatrix}$. Calculate the following:

1. $\mathbf{a} - \mathbf{b}$

2. $\mathbf{a} \cdot \mathbf{b}$

◇

Exercise 2.1 (Vector Algebra). Let $\mathbf{u} = \begin{pmatrix} 7 & 1 & -5 & 3 \end{pmatrix}$, $\mathbf{v} = \begin{pmatrix} 9 & -3 & 2 & 8 \end{pmatrix}$, $\mathbf{w} = \begin{pmatrix} 1 & 13 & -7 & 2 & 15 \end{pmatrix}$, and $c = 2$. Calculate the following:

1. $\mathbf{u} - \mathbf{v}$

2. $c\mathbf{w}$

3. $\mathbf{u} \cdot \mathbf{v}$

4. $\mathbf{w} \cdot \mathbf{v}$

◇

2.2 Linear Independence

2.2.1 Linear combinations

Definition 2.2 (Linear combinations). *The vector \mathbf{u} is a **linear combination** of the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ if*

$$\mathbf{u} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + \dots + c_k \mathbf{v}_k.$$



For example, $\begin{pmatrix} 9 & 13 & 17 \end{pmatrix}$ is a linear combination of the following three vectors: $\begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$, $\begin{pmatrix} 2 & 3 & 4 \end{pmatrix}$, and $\begin{pmatrix} 3 & 4 & 5 \end{pmatrix}$. This is because $\begin{pmatrix} 9 & 13 & 17 \end{pmatrix} = (2) \begin{pmatrix} 1 & 2 & 3 \end{pmatrix} + (-1) \begin{pmatrix} 2 & 3 & 4 \end{pmatrix} + (3) \begin{pmatrix} 3 & 4 & 5 \end{pmatrix}$.

2.2.2 Linear independence

Definition 2.3 (Linearly Independent Vectors). *A set of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ is linearly independent if the only solution to the equation*

$$c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + \dots + c_k \mathbf{v}_k = \mathbf{0}$$

is $c_1 = c_2 = \dots = c_k = 0$. If another solution exists, the set of vectors is linearly dependent.



A set S of vectors is linearly dependent *iff* at least one of the vectors in S can be written as a linear combination of the other vectors in S . Linear independence is only defined for sets of vectors with the same number of elements; any linearly independent set of vectors in n -space contains at most n vectors. Since $\begin{pmatrix} 9 & 13 & 17 \end{pmatrix}$ is a linear combination of $\begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$, $\begin{pmatrix} 2 & 3 & 4 \end{pmatrix}$, and $\begin{pmatrix} 3 & 4 & 5 \end{pmatrix}$, these 4 vectors constitute a linearly dependent set.

Example 2.2 (Linear Independence). Are the following sets of vectors linearly independent?

1. $\begin{pmatrix} 2 & 3 & 1 \end{pmatrix}$ and $\begin{pmatrix} 4 & 6 & 1 \end{pmatrix}$
2. $\begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$, $\begin{pmatrix} 0 & 5 & 0 \end{pmatrix}$, and $\begin{pmatrix} 10 & 10 & 0 \end{pmatrix}$



Exercise 2.2 (Linear Independence). Are the following sets of vectors linearly independent?

$$1. \mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \mathbf{v}_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$2. \mathbf{v}_1 = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} -4 \\ 6 \\ 5 \end{pmatrix}, \mathbf{v}_3 = \begin{pmatrix} -2 \\ 8 \\ 6 \end{pmatrix}$$

◇

2.3 Basics of Matrix Algebra

2.3.1 Matrix

Definition 2.4 (Matrix). A matrix is an array of real numbers arranged in m rows by n columns. The dimensionality of the matrix is defined as the number of rows by the number of columns, $m \times n$.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

♠

Note that you can think of vectors as special cases of matrices; a column vector of length k is a $k \times 1$ matrix, while a row vector of the same length is a $1 \times k$ matrix. It's also useful to think of matrices as being made up of a collection of row or column vectors. For example,

$$\mathbf{A} = (\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_m)$$

2.3.2 Matrix Addition

Let \mathbf{A} and \mathbf{B} be two $m \times n$ matrices.

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{pmatrix}$$

Note that matrices \mathbf{A} and \mathbf{B} must have the same dimensionality.

Example 2.3 (Matrix Addition). Given:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 1 & 2 \end{pmatrix}$$

Solve:

$$\mathbf{A} + \mathbf{B} =$$

◇

2.3.3 Scalar Multiplication

Given the scalar s , the scalar multiplication of $s\mathbf{A}$ is

$$s\mathbf{A} = s \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} = \begin{pmatrix} sa_{11} & sa_{12} & \cdots & sa_{1n} \\ sa_{21} & sa_{22} & \cdots & sa_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ sa_{m1} & sa_{m2} & \cdots & sa_{mn} \end{pmatrix}$$

Example 2.4 (Scalar Multiplication). Given:

$$s = 2, \quad \mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

Solve:

$$s\mathbf{A} =$$

◇

2.3.4 Matrix Multiplication

If \mathbf{A} is an $m \times k$ matrix and \mathbf{B} is a $k \times n$ matrix, then their product $\mathbf{C} = \mathbf{AB}$ is the $m \times n$ matrix where

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{ik}b_{kj}$$

We can restate the above equation for any combination of $n \times n$ (i.e. square) matrices, by simply multiplying the rows and columns.

Note that the number of columns of the first matrix must equal the number of rows of the second matrix. The sizes of the matrices (including the resulting product) must be

$$(m \times k)(k \times n) = (m \times n)$$

For example: Let the matrices \mathbf{A} and \mathbf{B} be defined as:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}.$$

Solve

$$\mathbf{AB} = \mathbf{C}.$$

First, check the dimensions of \mathbf{A} and \mathbf{B} , which are (2×3) and (3×2) , respectively. Using the rule above, we use the outer indices ($(\boxed{2} \times 3)$ and $(3 \times \boxed{2})$) and find that \mathbf{C} will be dimension (2×2) , e.g.

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}.$$

To solve for \mathbf{C} , matrix multiplication states that we use the dot product of the first *row* of \mathbf{A} by the first *column* of \mathbf{B} , etc., to get the result. We can use the notation

$$c_{11} = \mathbf{A}_{\text{row}_1} \cdot \mathbf{B}_{\text{col}_1}$$

$$c_{12} = \mathbf{A}_{\text{row}_1} \cdot \mathbf{B}_{\text{col}_2}$$

$$c_{21} = \mathbf{A}_{\text{row}_2} \cdot \mathbf{B}_{\text{col}_1}$$

$$c_{22} = \mathbf{A}_{\text{row}_2} \cdot \mathbf{B}_{\text{col}_2}$$

Or, more explicitly

$$c_{11} = a_{11} \times b_{11} + a_{12} \times b_{21} + a_{13} \times b_{31}$$

$$c_{12} = a_{11} \times b_{12} + a_{12} \times b_{22} + a_{13} \times b_{32}$$

$$c_{21} = a_{21} \times b_{11} + a_{22} \times b_{21} + a_{23} \times b_{31}$$

$$c_{22} = a_{21} \times b_{12} + a_{22} \times b_{22} + a_{23} \times b_{32}$$

The same steps used in the example above can be applied to any $n \times m$ or $n \times n$ matrix, provided that the rows and columns have an equal number of elements, e.g. the dimension of $\mathbf{A}_{\text{row}_i}$ is the same as $\mathbf{B}_{\text{col}_j}$.

Note that the matrix vector multiplication is just a simplification of the above example where \mathbf{B} would be a single column, replaced with the vector notation \mathbf{b} , such that

$$\mathbf{Ab} = \mathbf{c}$$

since \mathbf{c} is a vector.

Also note that if \mathbf{AB} exists, \mathbf{BA} exists only if $\dim(\mathbf{A}) = m \times n$ and $\dim(\mathbf{B}) = n \times m$. Generally $\mathbf{AB} \neq \mathbf{BA}$. Only in special circumstances is $\mathbf{AB} = \mathbf{BA}$ true, e.g. when \mathbf{A} or \mathbf{B} is the identity matrix \mathbf{I} , or $\mathbf{A} = \mathbf{B}^{-1}$.

Example 2.5 (Matrix Multiplication). Solve:

1. $\begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} =$
2. $\begin{pmatrix} 1 & 2 & -1 \\ 3 & 1 & 4 \end{pmatrix} \begin{pmatrix} -2 & 5 \\ 4 & -3 \\ 2 & 1 \end{pmatrix} =$

◇

2.3.5 Laws of Matrix Algebra

1. Associative: $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
 $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
2. Commutative: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
3. Distributive: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
 $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$

Note the order of multiplication matters:

$$\mathbf{AB} \neq \mathbf{BA}$$

For example,

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}$$

$$\mathbf{AB} = \begin{pmatrix} 2 & 3 \\ -2 & 2 \end{pmatrix}, \quad \mathbf{BA} = \begin{pmatrix} 1 & 7 \\ -1 & 3 \end{pmatrix}$$

2.3.6 Transpose

Definition 2.5 (Transpose). *The transpose of the $m \times n$ matrix \mathbf{A} is the $n \times m$ matrix \mathbf{A}^\top (also written \mathbf{A}' obtained by interchanging the rows and columns of \mathbf{A} .* ♠

Example 2.6 (Transpose).

$$\mathbf{A} = \begin{pmatrix} 4 & -2 & 3 \\ 0 & 5 & -1 \end{pmatrix}, \quad \mathbf{A}^\top = \begin{pmatrix} 4 & 0 \\ -2 & 5 \\ 3 & -1 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} 2 \\ -1 \\ 3 \end{pmatrix}, \quad \mathbf{B}^\top = (2 \quad -1 \quad 3)$$

◇

2.3.7 Properties of the transpose

The following rules apply for transposed matrices:

1. $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$
2. $(\mathbf{A}^\top)^\top = \mathbf{A}$
3. $(s\mathbf{A})^\top = s\mathbf{A}^\top$
4. $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$; and by induction $(\mathbf{ABC})^\top = \mathbf{C}^\top \mathbf{B}^\top \mathbf{A}^\top$

Example 2.7 (Matrix Multiplication). Given:

$$(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$$

It can be shown that

$$(\mathbf{AB})^\top = \left[\begin{pmatrix} 1 & 3 & 2 \\ 2 & -1 & 3 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 2 & 2 \\ 3 & -1 \end{pmatrix} \right]^\top = \begin{pmatrix} 12 & 7 \\ 5 & -3 \end{pmatrix}$$

and

$$\mathbf{B}^\top \mathbf{A}^\top = \begin{pmatrix} 0 & 2 & 3 \\ 1 & 2 & -1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & -1 \\ 2 & 3 \end{pmatrix} = \begin{pmatrix} 12 & 7 \\ 5 & -3 \end{pmatrix}.$$

◇

Exercise 2.3 (Matrix Multiplication). Given:

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & -1 & 1 \\ 1 & 2 & 0 & 1 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} 1 & 5 & -7 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \\ 2 & 0 & 0 \end{pmatrix}$$

$$\mathbf{C} = \begin{pmatrix} 3 & 2 & -1 \\ 0 & 4 & 6 \end{pmatrix}$$

Calculate the following:

1. \mathbf{AB}
2. \mathbf{BA}
3. $(\mathbf{BC})^\top$
4. \mathbf{BC}^\top

◇

2.4 Systems of Linear Equations

2.4.1 Linear Equation

The following equation is linear because there is only one variable per term and degree is at most 1

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b,$$

where a_i are parameters or coefficients, x_i are variables or unknowns.

We are often interested in solving linear systems like

$$\begin{array}{rclcl} x & - & 3y & = & -3 \\ 2x & + & y & = & 8 \end{array}$$

More generally, we might have a system of m equations in n unknowns

$$\begin{array}{cccccccl} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & & & \vdots & & \vdots & & \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array}$$

A **solution** to a linear system of m equations in n unknowns is a set of n numbers x_1, x_2, \dots, x_n that satisfy each of the m equations.

Example 2.8. $x = 3$ and $y = 2$ is the solution to the above 2×2 linear system. If you graph the two lines, you will find that they intersect at $(3, 2)$. \diamond

Does a linear system have one, no, or multiple solutions? For a system of 2 equations with 2 unknowns (i.e., two lines):

- **One solution:** The lines intersect at exactly one point.
- **No solution:** The lines are parallel.
- **Infinite solutions:** The lines coincide.

Methods to solve linear systems:

1. Substitution
2. Elimination of variables
3. Matrix methods

Exercise 2.4 (Linear Equations). Provide a system of 2 equations with 2 unknowns that has

1. one solution
2. no solution
3. infinite solutions

◇

2.5 Systems of Equations as Matrices

Matrices provide an easy and efficient way to represent linear systems such as

$$\begin{array}{cccccccl}
 a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\
 a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\
 \vdots & & & & \vdots & & & & \vdots \\
 a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m
 \end{array}$$

as

$$\mathbf{Ax} = \mathbf{b}.$$

2.5.1 Coefficient matrix

The $m \times n$ matrix \mathbf{A} is an array of m, n real numbers arranged in m rows by n columns:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

The unknown quantities are represented by the vector $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$.

The right hand side of the linear system is represented by the vector

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

2.5.2 Augmented Matrix

When we append \mathbf{b} to the coefficient matrix \mathbf{A} , we get the augmented matrix $\bar{\mathbf{A}} = [\mathbf{A}|\mathbf{b}]$:

$$\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array} \right)$$

Exercise 2.5 (Augmented Matrix). Create an augmented matrix that represent the following system of equations:

$$2x_1 - 7x_2 + 9x_3 - 4x_4 = 8$$

$$41x_2 + 9x_3 - 5x_6 = 11$$

$$x_1 - 15x_2 - 11x_5 = 9$$

◇

2.6 Solutions to Augmented Matrices & Systems of Equations

2.6.1 Row Echelon Form

Our goal is to translate our augmented matrix or system of equations into row echelon form. This will provide us with the values of the vector \mathbf{x} that solve the system. We use the row operations to change coefficients in the lower triangle of the augmented matrix to 0's. An augmented matrix of the form

$$\left(\begin{array}{cccc|c} \boxed{a'_{11}} & a'_{12} & a'_{13} & \cdots & a'_{1n} & b'_1 \\ 0 & \boxed{a'_{22}} & a'_{23} & \cdots & a'_{2n} & b'_2 \\ 0 & 0 & \boxed{a'_{33}} & \cdots & a'_{3n} & b'_3 \\ 0 & 0 & 0 & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \boxed{a'_{mn}} & b'_m \end{array} \right)$$

is said to be in *row echelon form*: each row has more leading zeros than the row preceding it.

2.6.2 Reduced Row Echelon Form

We can go one step further and put the matrix into *reduced row echelon form*. Reduced row echelon form makes the value of \mathbf{x} which solves the system very obvious. For a system of m equations in m unknowns, with no all-zero rows, the reduced row echelon form is

$$\left(\begin{array}{ccccc|c} \boxed{1} & 0 & 0 & 0 & 0 & b_1^* \\ 0 & \boxed{1} & 0 & 0 & 0 & b_2^* \\ 0 & 0 & \boxed{1} & 0 & 0 & b_3^* \\ 0 & 0 & 0 & \ddots & 0 & \vdots \\ 0 & 0 & 0 & 0 & \boxed{1} & b_m^* \end{array} \right)$$

2.6.3 Gaussian and Gauss-Jordan elimination

We can conduct elementary row operations to get our augmented matrix into row echelon or reduced row echelon form. The methods of transforming a matrix, or system, into row echelon and reduced row echelon form are referred to as Gaussian elimination and Gauss-Jordan elimination, respectively.

2.6.3.1 Elementary Row Operations

To do Gaussian and Gauss-Jordan elimination, we use three basic operations to transform the augmented matrix into another augmented matrix that represents an equivalent linear system: equivalent in the sense that the same values of x_j solve both the original and transformed matrix or system:

2.6.3.2 Interchanging Rows

Suppose we have the augmented matrix

$$\bar{\mathbf{A}} = \left(\begin{array}{cc|c} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \end{array} \right)$$

If we interchange the two rows, we get the augmented matrix

$$\left(\begin{array}{cc|c} a_{21} & a_{22} & b_2 \\ a_{11} & a_{12} & b_1 \end{array} \right)$$

which represents a linear system equivalent to that represented by matrix $\bar{\mathbf{A}}$.

2.6.3.3 Multiplying by a Constant

If we multiply the second row of matrix $\bar{\mathbf{A}}$ by a constant c , we get the augmented matrix

$$\left(\begin{array}{cc|c} a_{11} & a_{12} & b_1 \\ ca_{21} & ca_{22} & cb_2 \end{array} \right)$$

which represents a linear system equivalent to that represented by matrix $\bar{\mathbf{A}}$.

2.6.3.4 Adding (subtracting) Rows

If we add (subtract) the first row of matrix $\bar{\mathbf{A}}$ to (from) the second, we obtain the augmented matrix

$$\left(\begin{array}{cc|c} a_{11} & a_{12} & b_1 \\ a_{11} + a_{21} & a_{12} + a_{22} & b_1 + b_2 \end{array} \right)$$

which represents a linear system equivalent to that represented by matrix $\bar{\mathbf{A}}$.

Example 2.9. Solve the following system of equations by using elementary row operations:

$$\begin{array}{rclcl} x & - & 3y & = & -3 \\ 2x & + & y & = & 8 \end{array}$$

◇

Exercise 2.6 (Solving Systems of Equations). Put the following system of equations into augmented matrix form. Then, using Gaussian or Gauss-Jordan elimination, solve the system of equations by putting the matrix into row echelon or reduced row echelon form.

$$1. \begin{cases} x + y + 2z = 2 \\ 3x - 2y + z = 1 \\ y - z = 3 \end{cases}$$

$$2. \begin{cases} 2x + 3y - z = -8 \\ x + 2y - z = 12 \\ -x - 4y + z = -6 \end{cases}$$

◇

2.7 Rank & Number of Solutions

To determine how many solutions exist, we can use information about (1) the number of equations m , (2) the number of unknowns n , and (3) the **rank** of the matrix representing the linear system.

Definition 2.6 (Rank). *The maximum number of linearly independent row or column vectors in the matrix.* ♠

This is equivalent to the number of nonzero rows of a matrix in row echelon form. For any matrix \mathbf{A} , the row rank always equals column rank, and we refer to this number as the rank of \mathbf{A} .

For example $\begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{pmatrix}$ has Rank = 3, and $\begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 0 \end{pmatrix}$ has Rank = 2.

Exercise 2.7 (Rank of Matrices). Find the rank of each matrix below: (Hint: transform the matrices into row echelon form. Remember that the number of nonzero rows of a matrix in row echelon form is the rank of that matrix)

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

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

◇

2.8 The Inverse of a Matrix

Definition 2.7 (Identity Matrix). *The $n \times n$ identity matrix \mathbf{I}_n is the matrix whose diagonal elements are 1 and all off-diagonal elements are 0.* ♠

Examples:

$$\mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Definition 2.8 (Inverse Matrix). *An $n \times n$ matrix \mathbf{A} is nonsingular or invertible if there exists an $n \times n$ matrix \mathbf{A}^{-1} such that*

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$$

where \mathbf{A}^{-1} is the inverse of \mathbf{A} . If there is no such \mathbf{A}^{-1} , then \mathbf{A} is singular or not invertible. ♠

Example 2.10. Let

$$\mathbf{A} = \begin{pmatrix} 2 & 3 \\ 2 & 2 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} -1 & \frac{3}{2} \\ 1 & -1 \end{pmatrix}$$

Since

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$$

we conclude that \mathbf{B} is the inverse, \mathbf{A}^{-1} , of \mathbf{A} and that \mathbf{A} is nonsingular. \diamond

2.8.1 Properties of the Inverse

- If the inverse exists, it is unique.
- If \mathbf{A} is nonsingular, then \mathbf{A}^{-1} is nonsingular.
- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- If \mathbf{A} and \mathbf{B} are nonsingular, then \mathbf{AB} is nonsingular
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- If \mathbf{A} is nonsingular, then $(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top$

2.8.2 Procedure to Find the Inverse

Given \mathbf{A}^{-1} ; we know that if \mathbf{B} is the inverse of \mathbf{A} , then

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n.$$

Looking only at the first and last parts of this

$$\mathbf{AB} = \mathbf{I}_n.$$

Solving for \mathbf{B} is equivalent to solving for n linear systems, where each column of \mathbf{B} is solved for the corresponding column in \mathbf{I}_n . We can solve the systems simultaneously by augmenting \mathbf{A} with \mathbf{I}_n and performing Gauss-Jordan elimination on \mathbf{A} . If Gauss-Jordan elimination on $[\mathbf{A}|\mathbf{I}_n]$ results in $[\mathbf{I}_n|\mathbf{B}]$, then \mathbf{B} is the inverse of \mathbf{A} . Otherwise, \mathbf{A} is singular.

Therefore, to calculate the inverse of \mathbf{A} :

1. Form the augmented matrix $[\mathbf{A}|\mathbf{I}_n]$
2. Using elementary row operations, transform the augmented matrix to reduced row echelon form.
3. The result of step 2 is an augmented matrix $[\mathbf{C}|\mathbf{B}]$.

- (a) If $\mathbf{C} = \mathbf{I}_n$, then $\mathbf{B} = \mathbf{A}^{-1}$.
- (b) If $\mathbf{C} \neq \mathbf{I}_n$, then \mathbf{C} has a row of zeros. This means \mathbf{A} is singular and \mathbf{A}^{-1} does not exist.

Example 2.11. Find the inverse of the following matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 3 \\ 5 & 5 & 1 \end{pmatrix}$$

Solution. Solve using the following steps:

$$\begin{aligned} \mathbf{A}^{-1} &= \left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 2 & 3 & 0 & 1 & 0 \\ 5 & 5 & 1 & 0 & 0 & 1 \end{array} \right) \\ &= \left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 2 & 3 & 0 & 1 & 0 \\ 0 & 0 & -4 & -5 & 0 & 1 \end{array} \right) \\ &= \left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 2 & 3 & 0 & 1 & 0 \\ 0 & 0 & 1 & 5/4 & 0 & -1/4 \end{array} \right) \\ &= \left(\begin{array}{ccc|ccc} 1 & 1 & 0 & -1/4 & 0 & 1/4 \\ 0 & 2 & 0 & -15/4 & 1 & 3/4 \\ 0 & 0 & 1 & 5/4 & 0 & -1/4 \end{array} \right) \\ &= \left(\begin{array}{ccc|ccc} 1 & 1 & 0 & -1/4 & 0 & 1/4 \\ 0 & 1 & 0 & -15/8 & 1/2 & 3/8 \\ 0 & 0 & 1 & 5/4 & 0 & -1/4 \end{array} \right) \\ &= \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 13/8 & -1/2 & -1/8 \\ 0 & 1 & 0 & -15/8 & 1/2 & 3/8 \\ 0 & 0 & 1 & 5/4 & 0 & -1/4 \end{array} \right) \\ &= \left(\begin{array}{ccc|ccc} 13/8 & -1/2 & -1/8 \\ -15/8 & 1/2 & 3/8 \\ 5/4 & 0 & -1/4 \end{array} \right) \end{aligned}$$

◇

Exercise 2.8. Find the inverse of the following matrix

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

◇

2.9 Linear Systems and Inverses

Let's return to the matrix representation of a linear system

$$\mathbf{Ax} = \mathbf{b}$$

If \mathbf{A} is an $n \times n$ matrix, then $\mathbf{Ax} = \mathbf{b}$ is a system of n equations in n unknowns. Suppose \mathbf{A} is nonsingular. Then \mathbf{A}^{-1} exists. To solve this system, we can multiply each side by \mathbf{A}^{-1} and reduce it as follows:

$$\begin{aligned}\mathbf{A}^{-1}(\mathbf{Ax}) &= \mathbf{A}^{-1}\mathbf{b} \\ (\mathbf{A}^{-1}\mathbf{A})\mathbf{x} &= \mathbf{A}^{-1}\mathbf{b} \\ \mathbf{I}_n\mathbf{x} &= \mathbf{A}^{-1}\mathbf{b} \\ \mathbf{x} &= \mathbf{A}^{-1}\mathbf{b}\end{aligned}$$

Hence, given \mathbf{A} and \mathbf{b} , and given that \mathbf{A} is nonsingular, then $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ is a unique solution to this system.


Exercise 2.9 (Solve linear system using inverses). Use the inverse matrix to solve the following linear system:

$$\begin{aligned}-3x + 4y &= 5 \\ 2x - y &= -10\end{aligned}$$

Hint: the linear system above can be written in the matrix form $\mathbf{Az} = \mathbf{b}$ given

$$\mathbf{A} = \begin{pmatrix} -3 & 4 \\ 2 & -1 \end{pmatrix}, \mathbf{z} = \begin{pmatrix} x \\ y \end{pmatrix}, \text{ and } \mathbf{b} = \begin{pmatrix} 5 \\ -10 \end{pmatrix}. \quad \diamond$$

2.10 Determinants

Definition 2.9 (Nonsingular). A square matrix is nonsingular iff its determinant is not zero. 

Determinants can be used to determine whether a square matrix is nonsingular. The determinant of a 1×1 matrix, \mathbf{A} , equals a_{11} . The determinant of a 2×2 matrix,

$$\mathbf{A} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

is

$$\begin{aligned}\det(\mathbf{A}) &= |\mathbf{A}| \\ &= a_{11}|a_{22}| - a_{12}|a_{21}| \\ &= a_{11}a_{22} - a_{12}a_{21}\end{aligned}$$

We can extend the second to last equation above to get the definition of the determinant of a 3×3 matrix:

$$\begin{aligned} \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} &= a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) \\ &\quad - a_{12}(a_{21}a_{33} - a_{23}a_{31}) \\ &\quad + a_{13}(a_{21}a_{32} - a_{22}a_{31}) \end{aligned}$$

Let's extend this now to any $n \times n$ matrix. Let's define \mathbf{A}_{ij} as the $(n-1) \times (n-1)$ sub-matrix of \mathbf{A} obtained by deleting row i and column j . Let the (i, j) -th **minor** of \mathbf{A} be the determinant of \mathbf{A}_{ij} :

$$M_{ij} = |\mathbf{A}_{ij}|$$

Then for any $n \times n$ matrix \mathbf{A}

$$|\mathbf{A}| = a_{11}M_{11} - a_{12}M_{12} + \cdots + (-1)^{n+1}a_{1n}M_{1n}$$

Example 2.12 (Determinants). Does the following matrix have an inverse?

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 3 \\ 5 & 5 & 1 \end{pmatrix}$$

We find the solution by:

1. Calculate its determinant.

$$\begin{aligned} &= 1(2 - 15) - 1(0 - 15) + 1(0 - 10) \\ &= -13 + 15 - 10 \\ &= -8 \end{aligned}$$

2. Since $|\mathbf{A}| \neq 0$, we conclude that \mathbf{A} has an inverse.

◇

Exercise 2.10 (Determinants and Inverses). Determine whether the following matrices are nonsingular:

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

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

◇

2.11 Matrix Inverse using the Determinant

Thus far, we have algorithms to

1. Find the solution of a linear system,
2. Find the inverse of a matrix.

At this point, we have no way of telling how the solutions x_j change as the parameters a_{ij} and b_i change, except by changing the values and “re-solving” the algorithms.

With determinants, we have an explicit formula for the inverse, and therefore an explicit formula for the solution of an $n \times n$ linear system. Hence, we can examine how changes in the parameters and b_i affect the solutions x_j .

Definition 2.10 (Determinant Formula for the Inverse). *The determinant of a 2×2 matrix $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is defined as:*

$$\frac{1}{\det(\mathbf{A})} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$



Example 2.13 (Determinants and Inverses). Calculate the inverse of matrix \mathbf{A} from Exercise 2.9 using the determinant formula. Recall,

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

then:

$$\det(\mathbf{A}) = (-3)(-1) - (4)(2) = 3 - 8 = -5$$

$$\frac{1}{\det(\mathbf{A})} \begin{pmatrix} -1 & -4 \\ -2 & -3 \end{pmatrix}$$

$$\frac{1}{-5} \begin{pmatrix} -1 & -4 \\ -2 & -3 \end{pmatrix}$$

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



Exercise 2.11 (Calculate Inverse using Determinant Formula). Calculate the inverse of \mathbf{A} , where

$$\mathbf{A} = \begin{pmatrix} 3 & 5 \\ -7 & 2 \end{pmatrix}.$$



Chapter 3

Functions and Operations

3.1 Summation and Product Operators

Addition (+), Subtraction (-), multiplication and division are basic operations of arithmetic – combining numbers. In statistics and calculus, we want to add a *sequence* of numbers that can be expressed as a pattern without needing to write down all its components.

For example, how would we express the sum of all numbers from 1 to 100 without writing a hundred numbers? For this we use the summation operator \sum and the product operator \prod .

3.1.1 Summation

Define the sum of x_i from 1 to n as

$$\sum_{i=1}^n x_i = x_1 + x_2 + x_3 + \cdots + x_n$$

The bottom of the \sum symbol indicates the index i you are summing and its starting value; here, $i = 1$. The top value is where the index ends. The notion of “addition” is part of the \sum symbol. The content to the right of the summation is what we add. While you can pick any index, start, and end values, the content on the right must also have the index.

Summation Properties:

- $$\sum_{i=1}^n cx_i = c \sum_{i=1}^n x_i$$

- $\sum_{i=1}^n (x_i + y_i) = \sum_{i=1}^n x_i + \sum_{i=1}^n y_i$
- $\sum_{i=1}^n c = nc$

3.1.2 Product

Define the product of x_i from 1 to n as

$$\prod_{i=1}^n x_i = x_1 \times x_2 \times x_3 \times \cdots \times x_n$$

The bottom of the \prod symbol indicates the index i you are multiplying and its starting value; here, $i = 1$. The top value is where the index ends. The notion of “multiplication” is part of the \prod symbol.

Product Properties:

- $\prod_{i=1}^n cx_i = c^n \prod_{i=1}^n x_i$
- $\prod_{i=k}^n cx_k = c^{n-k} \prod_{i=k}^n x_i$
- $\prod_{i=1}^n c = c^n$

3.1.3 Factorials!

The symbol $!$ is called a *factorial*, and we use it to find the factorial of a variable or function. For example the factorial of x is

$$x! = x \cdot (x-1) \cdot (x-2) \cdots (1)$$

or, for example where $x = 5$

$$\begin{aligned} 5! &= 5 \cdot (5-1) \cdot (5-2) \cdot (5-3) \cdot (5-4) \\ &= 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \\ &= 120. \end{aligned}$$

3.1.4 Modulo

Modulo tells you the remainder when you divide the first number by the second. For example

$$100 \% 30 = 10$$

could also be written as

$$100 \bmod 30 = 10.$$

Example 3.1 (Operators). Lets look at a few examples

1. $\sum_{i=1}^5 i =$

2. $\prod_{i=1}^5 i =$

3. $14 \bmod 4 =$

4. $4! =$

◇

Exercise 3.1 (Operators). Let $x_1 = 4, x_2 = 3, x_3 = 7, x_4 = 11, x_5 = 2$. Solve the following:

1. $\sum_{i=1}^3 (7)x_i$

2. $\sum_{i=1}^5 2$

3. $\prod_{i=3}^5 (2)x_i$

◇

3.2 Introduction to Functions

3.2.1 Functions

A **function** (in \mathbb{R}^1) is a mapping, or transformation, that relates members of one set to members of another set. For instance, if you have two sets: set A and set B , a function from A to B maps every value a in set A such that $f(a) \in B$. Functions can be “many-to-one”, where many values or combinations of values from set A produce a single output in set B , or they can be “one-to-one”, where each value in set A corresponds to a single value in set B . A function by definition has a single function value for each element of its domain. This means, there cannot be “one-to-many” mapping.

3.2.2 Dimensionality

\mathbb{R}^1 is the set of all real numbers extending from $-\infty$ to $+\infty$, i.e., the real number line. \mathbb{R}^n is an n -dimensional space, where each of the n axes extends from $-\infty$ to $+\infty$.

- \mathbb{R}^1 is a one dimensional line.
- \mathbb{R}^2 is a two dimensional plane.
- \mathbb{R}^3 is a three dimensional space.

Points in \mathbb{R}^n are ordered n -tuples (just means an combination of n elements where order matters), where each element of the n -tuple represents the coordinate along that dimension.

For example:

- \mathbb{R}^1 : (3)
- \mathbb{R}^2 : (-15, 5)
- \mathbb{R}^3 : (86, 4, 0)

Example: Function of one variable. $f : \mathbb{R}^1 \rightarrow \mathbb{R}^1$

- $f(x) = x + 1$. For each x in \mathbb{R}^1 , $f(x)$ assigns the number $x + 1$.

Example: Function of two variables. $f : \mathbb{R}^2 \rightarrow \mathbb{R}^1$.

- $f(x, y) = x^2 + y^2$. For each ordered pair (x, y) in \mathbb{R}^2 , $f(x, y)$ assigns the number $x^2 + y^2$.

We often use variable x as input and another y as output, e.g. $y = x + 1$

Example 3.2 (Functions). For each of the following, state whether they are one-to-one or many-to-one functions.

1. For $x \in [0, \infty]$, $f : x \rightarrow x^2$ (this could also be written as $f(x) = x^2$).
2. For $x \in [-\infty, \infty]$, $f : x \rightarrow x^2$.

◇

Exercise 3.2 (Functions). For each of the following, state whether they are one-to-one or many-to-one functions.

1. For $x \in [-3, \infty]$, $f : x \rightarrow x^2$.
2. For $x \in [0, \infty]$, $f : x \rightarrow \sqrt{x}$.

◇

3.2.3 Definitions

Some functions are defined only on proper subsets of \mathbb{R}^n .

- **Domain:** the set of numbers in X at which $f(x)$ is defined.
- **Range:** elements of Y assigned by $f(x)$ to elements of X , or

$$f(X) = \{y : y = f(x), x \in X\}.$$

Most often used when talking about a function $f : \mathbf{R}^1 \rightarrow \mathbf{R}^1$.

- **Image:** same as range, but more often used when talking about a function $f : \mathbf{R}^n \rightarrow \mathbf{R}^1$.

Some General Types of Functions

- **Monomials:** $f(x) = ax^k$, where a is the coefficient, k is the degree. Examples: $y = x^2$, $y = -\frac{1}{2}x^3$.
- **Polynomials:** sum of monomials. Examples: $y = -\frac{1}{2}x^3 + x^2$, $y = 3x + 5$. The degree of a polynomial is the highest degree of its monomial terms. Also, it's often a good idea to write polynomials with terms in decreasing degree.
- **Exponential Functions:** Example: $y = 2^x$.

3.3 log and exponent

3.3.1 Relationship of logarithmic and exponential functions

Given:

$$y = \log_a(x) \iff a^y = x$$

The log function can be thought of as an inverse for exponential functions. a is referred to as the “base” of the logarithm.

3.3.2 Common Bases

The two most common logarithms are base 10 and base e .

1. Base 10: $y = \log_{10}(x) \iff 10^y = x$. The base 10 logarithm is often simply written as “ $\log(x)$ ” with no base denoted.
2. Base e : $y = \log_e(x) \iff e^y = x$. The base e logarithm is referred to as the “natural” logarithm and is written as “ $\ln(x)$ ”.

3.3.3 Properties of exponential functions

- $a^x a^y = a^{x+y}$
- $a^{-x} = 1/a^x$
- $a^x / a^y = a^{x-y}$
- $(a^x)^y = a^{xy}$
- $a^0 = 1$

3.3.4 Properties of logarithmic functions of any base

Generally, when statisticians or data scientists write $\log(x)$ they mean $\log_e(x)$. In other words $\log_e(x) \equiv \ln(x) \equiv \log(x)$. Therefore,

$$\log_a(a^x) = x$$

and

$$a^{\log_a(x)} = x$$

- $\log(xy) = \log(x) + \log(y)$
- $\log(x^y) = y \log(x)$
- $\log(1/x) = \log(x^{-1}) = -\log(x)$
- $\log(x/y) = \log(x \cdot y^{-1}) = \log(x) + \log(y^{-1}) = \log(x) - \log(y)$
- $\log(1) = \log(e^0) = 0$

3.3.5 Change of Base Formula

Use the change of base formula to switch bases as necessary:

$$\log_b(x) = \frac{\log_a(x)}{\log_a(b)}$$

Example:

$$\log_{10}(x) = \frac{\ln(x)}{\ln(10)}$$

3.3.6 Log, Product, and Sum Operators

You can use logs to go between sum and product notation. This will be particularly important when you're learning maximum likelihood estimation.

$$\begin{aligned}\log\left(\prod_{i=1}^n x_i\right) &= \log(x_1 \cdot x_2 \cdot x_3 \cdots x_n) \\ &= \log(x_1) + \log(x_2) + \log(x_3) + \cdots + \log(x_n) \\ &= \sum_{i=1}^n \log(x_i)\end{aligned}$$

Therefore, you can see that the log of a product is equal to the sum of the logs. We can write this more generally by adding in a constant, c :

$$\begin{aligned}\log\left(\prod_{i=1}^n cx_i\right) &= \log(cx_1 \cdot cx_2 \cdots cx_n) \\ &= \log(c^n \cdot x_1 \cdot x_2 \cdots x_n) \\ &= \log(c^n) + \log(x_1) + \log(x_2) + \cdots + \log(x_n) \\ &= n \log(c) + \sum_{i=1}^n \log(x_i)\end{aligned}$$

Example 3.3 (Logarithmic Functions). Evaluate each of the following logarithms

1. $\log_4(16)$
2. $\log_2(16)$
3. $\log_4(x^3y^5)$, Hint: Simplify using as many of the logarithmic properties as you can.

◇

Exercise 3.3 (Logarithmic Functions). Evaluate each of the following logarithms. Simplify using as many of the logarithmic properties as you can.

1. $\log_{\frac{3}{2}}\left(\frac{27}{8}\right)$
2. $\log\left(\frac{x^9y^5}{z^3}\right)$

3. $\ln \sqrt{xy}$

◇

3.4 Graphing Functions

What can a graph tell you about a function?

- Is the function increasing or decreasing? Over what part of the domain?
- How “fast” does it increase or decrease?
- Are there global or local maxima and minima? Where?
- Are there inflection points?
- Is the function continuous?
- Is the function differentiable?
- Does the function tend to some limit?
- Other questions related to the substance of the problem at hand.

3.5 Solving for Variables and Finding Roots

Sometimes we’re given a function $y = f(x)$ and we want to find how x varies as a function of y . Use algebra to move x to the left hand side (LHS) of the equation and so that the right hand side (RHS) is only a function of y .

Example 3.4 (Solving for Variables). Solve for x :

1. $y = 3x + 2$

2. $y = e^x$

◇

Solving for variables is especially important when we want to find the **roots** of an equation: those values of variables that cause an equation to equal zero. Especially important in finding equilibria and in doing maximum likelihood estimation.

3.5.1 Procedure

Given $y = f(x)$, set $f(x) = 0$. Solve for x .

Example with multiple roots:

$$\begin{aligned}f(x) = x^2 - 9 &\implies 0 = x^2 - 9 \\&\implies 9 = x^2 \\&\implies \pm\sqrt{9} = \sqrt{x^2} \\&\implies \pm 3 = x\end{aligned}$$

3.5.2 Quadratic Formula

For quadratic equations $ax^2 + bx + c = 0$, use the quadratic formula:

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

Exercise 3.4 (Finding Roots). Solve for x :

1. $f(x) = 3x + 2 = 0$
2. $f(x) = x^2 + 3x - 4 = 0$
3. $f(x) = e^{-x} - 10 = 0$

◇

3.6 Sets

3.6.1 Interior Point

The point \mathbf{x} is an interior point of the set S if \mathbf{x} is in S and if there is some ϵ -ball around \mathbf{x} that contains only points in S . The **interior** of S is the collection of all interior points in S . The interior can also be defined as the union of all open sets in S .

- If the set S is circular, the interior points are everything inside of the circle, but not on the circle's rim.
- Example: The interior of the set $\{(x, y) : x^2 + y^2 \leq 4\}$ is $\{(x, y) : x^2 + y^2 < 4\}$.

3.6.2 Boundary Point

The point \mathbf{x} is a boundary point of the set S if every ϵ -ball around \mathbf{x} contains both points that are in S and points that are outside S . The **boundary** is the collection of all boundary points.

- If the set S is circular, the boundary points are everything on the circle's rim.
- Example: The boundary of $\{(x, y) : x^2 + y^2 \leq 4\}$ is $\{(x, y) : x^2 + y^2 = 4\}$.

3.6.3 Open

A set S is open if for each point \mathbf{x} in S , there exists an open ϵ -ball around \mathbf{x} completely contained in S .

- If the set S is circular and open, the points contained within the set get infinitely close to the circle's rim, but do not touch it.
- Example: $\{(x, y) : x^2 + y^2 < 4\}$.

3.6.4 Closed

A set S is closed if it contains all of its boundary points.

- Alternatively: A set is closed if its complement is open.
- If the set S is circular and closed, the set contains all points within the rim as well as the rim itself.
- Example: $\{(x, y) : x^2 + y^2 \leq 4\}$
- Note: a set may be neither open nor closed. Example: $\{(x, y) : 2 < x^2 + y^2 \leq 4\}$.

3.6.5 Complement

The complement of set S is everything outside of S .

- If the set S is circular, the complement of S is everything outside of the circle.
- Example: The complement of $\{(x, y) : x^2 + y^2 \leq 4\}$ is $\{(x, y) : x^2 + y^2 > 4\}$.

3.6.6 Empty

The empty (or null) set is a unique set that has no elements, denoted by $\{\}$ or \emptyset .

- The empty set is an example of a set that is open and closed, or a “clopen” set.
- Examples: The set of squares with 5 sides; the set of countries south of the South Pole.

Chapter 4

Limits

Solving limits, i.e. finding out the value of functions as the input moves closer to some value. This is important for the data scientist's mathematical toolkit for two related tasks. First is the study of calculus, which will be in turn useful to show where certain functions are maximized or minimized. The second is for the study of statistical inference, which is the study of inferring something about things you cannot see, by using something you can see.

4.1 The Central Limit Theorem

Perhaps the most important theorem in statistics is the Central Limit Theorem,

Theorem 4.1 (Central Limit Theorem (i.i.d. case)). *For any series of independent and identically distributed random variables X_1, X_2, \dots , we know the distribution of its sum, even if we do not know the distribution of X . The distribution of the sum is a Normal distribution:*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \text{Normal}(0, 1),$$

where μ is the mean of X and σ is the standard deviation of X . The arrow is read as “converges in distribution to”. $\text{Normal}(0, 1)$ indicates a Normal Distribution with mean 0 and variance 1.

That is, the limit of the distribution of the LHS is the distribution of the RHS side. ♠

The sign of a limit is the arrow “ \rightarrow ”. Although we have not yet covered probability (Section 7), and therefore have not described what

distributions and random variables are, it is mentioning the Central Limit Theorem. The Central Limit Theorem is powerful because it gives us a *guarantee* of what would happen if $n \rightarrow \infty$, which in this case means we collected more data.

4.2 The Law of Large Numbers

A rival to the Central Limit Theorem, is the Law of Large Numbers:

Theorem 4.2 ((Weak) Law of Large Numbers). *For any draw of identically distributed independent variables with mean μ , the sample average after n draws, \bar{X}_n , converges in probability to the true mean as $n \rightarrow \infty$:*

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

Simplified is $\bar{X}_n \xrightarrow{P} \mu$, where the arrow is read as “converges in probability to”. ♠

Intuitively, the more data, the more accurate is your guess. For example, the Figure 4.1 shows how the sample average from many coin tosses converges to the true value : 0.5.

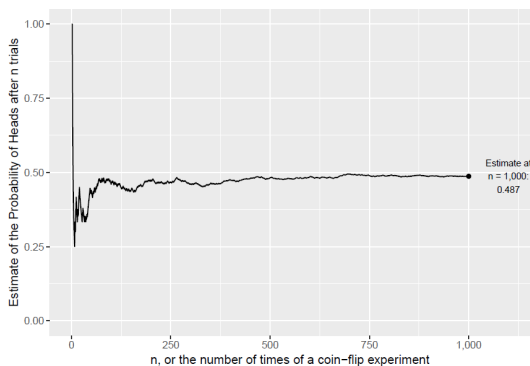


Figure 4.1: As the number of coin tosses goes to infinity, the average probability of heads converges to 0.5

4.3 Sequences

We need a couple of steps until we get to limit theorems in probability. First we will introduce a “sequence”, then we will think about the limit of a sequence, then we will think about the limit of a *function*.

A sequence

$$\{x_n\} = \{x_1, x_2, x_3, \dots, x_n\}$$

is an ordered set of real numbers, where x_1 is the first term in the sequence and y_n is the n -th term. Generally, a sequence is infinite, that is $n = \infty$. We can also write the sequence as

$$\{x_n\}_{n=1}^{\infty}$$

where the subscript and superscript are read together as “from 1 to infinity.”

Example 4.1 (Sequences). How does these sequence behave?

1. $\{A_n\} = \left\{2 - \frac{1}{n^2}\right\}$
2. $\{B_n\} = \left\{\frac{n^2+1}{n}\right\}$
3. $\{C_n\} = \left\{(-1)^n \left(1 - \frac{1}{n}\right)\right\}$

◇

We find the sequence by simply “plugging in” the integers into each n . The important thing is to get a sense of how these numbers are going to change. Example 1’s numbers seem to come closer and closer to 2, but will it ever exceed 2? Example 2’s numbers are also increasing each time, but will it hit a limit? What is the pattern in Example 3? Graphing helps you make this point more clearly. See the sequence of $n = 1, \dots, 20$ for each of the three examples in Figure 4.2.

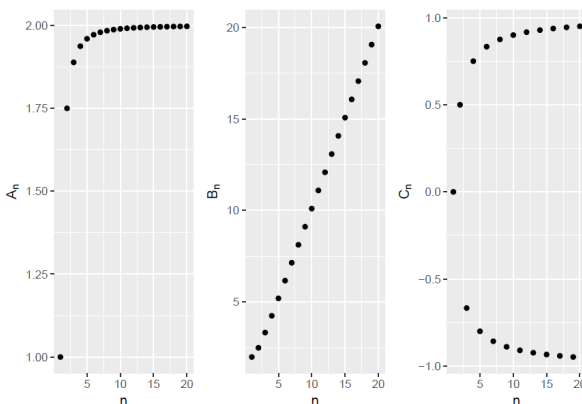


Figure 4.2: Behavior of Example 4.1 Sequences.

4.4 The Limit of a Sequence

The notion of “converging to a limit” is the behavior of the points in Example 4.1. In some sense, that’s the counter-factual we want to know. What happens as $n \rightarrow \infty$?

1. Sequences like 4.1-1 above that converge to a limit.
2. Sequences like 4.1-2 above that increase without bound.
3. Sequences like 4.1-3 above that neither converge nor increase without bound — alternating over the number line.

Definition 4.1 (The Limit of a Sequence). *The sequence $\{y_n\}$ has the limit L , which we write as*

$$\lim_{n \rightarrow \infty} y_n = L.$$

If for any $\epsilon > 0$ there is an integer N (which depends on ϵ) with the property that $|y_n - L| < \epsilon$ for each $n > N$, $\{y_n\}$ is said to converge to L . If the above does not hold, then $\{y_n\}$ diverges.



We can also express the behavior of a sequence as bounded or not:

1. Bounded: if $|y_n| \leq K$ for all n .
2. Monotonically Increasing: $y_{n+1} > y_n$ for all n .
3. Monotonically Decreasing: $y_{n+1} < y_n$ for all n .

A limit is *unique*: If $\{y_n\}$ converges, then the limit L is unique.

If a sequence converges, then the sum of such sequences also converges. Let $\lim_{n \rightarrow \infty} y_n = y$ and $\lim_{n \rightarrow \infty} z_n = z$. Then

1. $\lim_{n \rightarrow \infty} [ky_n + \ell z_n] = ky + \ell z$.
2. $\lim_{n \rightarrow \infty} y_n z_n = yz$.
3. $\lim_{n \rightarrow \infty} \frac{y_n}{z_n} = \frac{y}{z}$, provided $z \neq 0$.

This looks reasonable enough. So when *don’t* the parts of the fraction converge? If $\lim_{n \rightarrow \infty} y_n = \infty$ and $\lim_{n \rightarrow \infty} z_n = \infty$, what is $\lim_{n \rightarrow \infty} y_n - z_n$? What is $\lim_{n \rightarrow \infty} \frac{y_n}{z_n}$?

It is nice for a sequence to converge in limit. We want to know if complex-looking sequences converge or not. The solution is to break

that complex sequence up into sums of simple fractions where n only appears in the denominator: $\frac{1}{n}$, $\frac{1}{n^2}$, etc. Each of these will converge to 0, because the denominator continues to increase. Then, because of the properties above, we can then find the final sequence.

Example 4.2 (Simplifying a Fraction into Sums). Find the limit of

$$\lim_{n \rightarrow \infty} \frac{n+3}{n}.$$

At first glance, $n+3$ and n both grow to ∞ , so it looks like we need to divide infinity by infinity. However, we can express this fraction as a sum, then the limits apply separately:

$$\lim_{n \rightarrow \infty} \frac{n+3}{n} = \lim_{n \rightarrow \infty} \left(1 + \frac{3}{n} \right) = \underbrace{\lim_{n \rightarrow \infty} 1}_1 + \underbrace{\lim_{n \rightarrow \infty} \left(\frac{3}{n} \right)}_0$$

so, the limit is actually 1. ◇

After some practice, the key to intuition is whether one part of the fraction grows “faster” than another. If the denominator grows faster to infinity than the numerator, then the fraction will converge to 0, even if the numerator also increases to infinity. In a sense, limits show that all infinities are not the same.

Exercise 4.3. Find the following limits of sequences, then explain in English the intuition for why that is the case.

1. $\lim_{n \rightarrow \infty} \frac{2n}{n^2+1}$
2. $\lim_{n \rightarrow \infty} (n^3 - 100n^2)$

◇

4.5 Limits of a Function

We’ve now covered functions and just covered limits of sequences, so now is the time to combine the two.

A function f is a compact representation of some behavior we care about. Like for sequences, we often want to know if $f(x)$ approaches some number L as its independent variable x moves to some number c (which is usually 0 or $\pm\infty$). If it does, we say that the limit of $f(x)$, as x approaches c , is L : $\lim_{x \rightarrow c} f(x) = L$. Unlike a sequence, x is a continuous number, and we can move in decreasing order as well as increasing.

For a limit L to exist, the function $f(x)$ must approach L from both the left (increasing) and the right (decreasing).

Definition 4.2 (Limit of a function). Let $f(x)$ be defined at each point in some open interval containing the point c . Then L equals $\lim_{x \rightarrow c} f(x)$ if for any (small positive) number ϵ , there exists a corresponding number $\delta > 0$ such that if $0 < |x - c| < \delta$, then $|f(x) - L| < \epsilon$.



A nice, if subtle result, is that $f(x)$ does not necessarily have to be defined at c for $\lim_{x \rightarrow c}$ to exist.

4.5.1 Properties of Limits

Let f and g be functions with $\lim_{x \rightarrow c} f(x) = k$ and $\lim_{x \rightarrow c} g(x) = \ell$.

1. $\lim_{x \rightarrow c} [f(x) + g(x)] = \lim_{x \rightarrow c} f(x) + \lim_{x \rightarrow c} g(x)$.
2. $\lim_{x \rightarrow c} kf(x) = k \lim_{x \rightarrow c} f(x)$.
3. $\lim_{x \rightarrow c} f(x)g(x) = \left[\lim_{x \rightarrow c} f(x) \right] \cdot \left[\lim_{x \rightarrow c} g(x) \right]$.
4. $\lim_{x \rightarrow c} \frac{f(x)}{g(x)} = \frac{\lim_{x \rightarrow c} f(x)}{\lim_{x \rightarrow c} g(x)}$, provided $\lim_{x \rightarrow c} g(x) \neq 0$.

Simple limits of functions can be solved as we did limits of sequences. Just be careful which part of the function is changing. Limits can get more complex in roughly two ways. First, the functions may become large polynomials with many moving pieces. Second, the functions may become discontinuous.

There are a few more characteristics of limits:

1. Right-hand limit: The value approached by $f(x)$ when you move from right to left.
2. Left-hand limit: The value approached by $f(x)$ when you move from left to right.
3. Infinity: The value approached by $f(x)$ as x grows infinitely large. Sometimes this may be a number; sometimes it might be ∞ or $-\infty$.
4. Negative infinity: The value approached by $f(x)$ as x grows infinitely negative. Sometimes this may be a number; sometimes it might be ∞ or $-\infty$.

The distinction between left and right becomes important when the function is not determined for some values of x . What are those cases in the examples below?

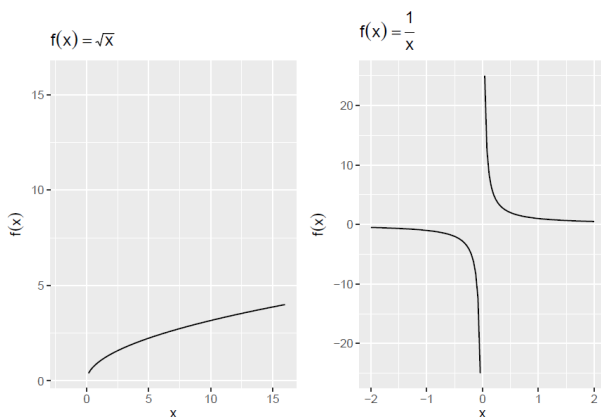


Figure 4.3: Functions which are not defined in some areas

Example 4.3 (Limits of Functions). Find the limit of the following functions.

1. $\lim_{x \rightarrow c} k$

2. $\lim_{x \rightarrow c} x$

3. $\lim_{x \rightarrow 2} (2x - 3)$

4. $\lim_{x \rightarrow c} x^n$

◇

The function can be thought of as a more general or “smooth” version of sequences. For example,

Exercise 4.4 (Limits of a Fraction of Functions). Find the limit of

$$\lim_{x \rightarrow \infty} \frac{(x^4 + 3x - 99)(2 - x^5)}{(18x^7 + 9x^6 - 3x^2 - 1)(x + 1)}$$

◇

Exercise 4.5. Solve the following limits of functions

1. $\lim_{x \rightarrow 0} |x|$

2. $\lim_{x \rightarrow 0} \left(1 + \frac{1}{x^2}\right)$

◇

4.6 Continuity

To repeat a finding from the limits of functions: $f(x)$ does not necessarily have to be defined at c for $\lim_{x \rightarrow c}$ to exist. Functions that have breaks in their lines are called discontinuous. Functions that have no breaks are called continuous. Continuity is a concept that is more fundamental, but related, to that of “differentiability”, which we will cover in calculus.

Definition 4.3 (Continuity). *Suppose that the domain of the function f includes an open interval containing the point c . Then f is continuous at c if $\lim_{x \rightarrow c} f(x)$ exists and if $\lim_{x \rightarrow c} f(x) = f(c)$. Further, f is continuous on an open interval (a, b) if it is continuous at each point in the interval.*



To prove that a function is continuous for all points is beyond this practical introduction to math, but the general intuition can be grasped by graphing.

Example 4.4 (Continuous and Discontinuous Functions). For each function, determine if it is continuous or discontinuous.

1. $f(x) = \sqrt{x}$.
2. $f(x) = e^x$.
3. $f(x) = 1 + \frac{1}{x^2}$.
4. $f(x) = \text{floor}(x)$.

The floor is the smaller of the two integers bounding a number. So $\text{floor}(x = 2.999) = 2$, $\text{floor}(x = 2.0001) = 2$, and $\text{floor}(x = 2) = 2$.

Solution. In Figure 4.4, we can see that the first two functions are continuous, and the next two are discontinuous. $f(x) = 1 + \frac{1}{x^2}$ is discontinuous at $x = 0$, and $f(x) = \text{floor}(x)$ is discontinuous at each whole number. \diamond

4.6.1 Properties of Continuous Functions

1. **Continuous:** If f and g are continuous at point c , then $f + g$, $f - g$, $f \cdot g$, $|f|$, and αf are continuous at point c also. f/g is continuous, provided $g(c) \neq 0$.
2. **Boundedness:** If f is continuous on the closed bounded interval $[a, b]$, then there is a number K such that $|f(x)| \leq K$ for each x in $[a, b]$.

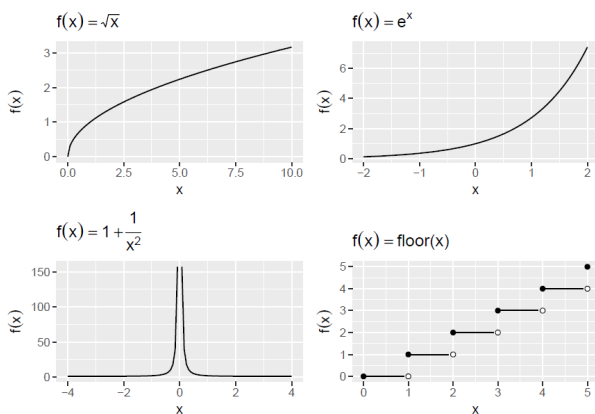


Figure 4.4: Continuous and Discontinuous Functions

3. Max/Min: If f is continuous on the closed bounded interval $[a, b]$, then f has a maximum and a minimum on $[a, b]$. They may be located at the end points.

Exercise 4.6 (Limit when Denominator converges to 0). Let

$$f(x) = \frac{x^2 + 2x}{x}.$$

1. Graph the function. Is it defined everywhere?
2. What is the functions limit at $x \rightarrow 0$?

◇

Chapter 5

Calculus

Calculus is a fundamental part of any type of statistics exercise. Although you may not be taking derivatives and integral in your daily work as an analyst, calculus undergirds many concepts we use: maximization, expectation, and cumulative probability.

5.1 The Mean is a Type of Integral

The average of a quantity is a type of weighted mean, where the potential values are weighted by their likelihood, loosely speaking. The integral is actually a general way to describe this weighted average when there are conceptually an infinite number of potential values.

If X is a continuous random variable, its expected value $E(X)$, i.e. the center of mass, is given by

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

where $f(x)$ is the probability density function of X .

If X is a discrete random variable, its expected value $E(X)$ is

$$E(X) = \sum_{j=1}^{\infty} x_j P(X = x_j)$$

Even more concretely, if the potential values of X are finite, then we can write out the expected value as a weighted mean, where the weights is the probability that the value occurs.

$$E(X) = \sum_x \left(\underbrace{x}_{\text{value}} \cdot \underbrace{P(X = x)}_{\text{weight, or PMF}} \right)$$

5.2 Derivatives

The derivative of f at x is its rate of change at x : how much $f(x)$ changes with a change in x . The rate of change is a fraction — rise over run — but because not all lines are straight and the rise over run formula will give us different values depending on the range we examine, we need to take a limit (see Section 4).

Definition 5.1 (Derivative). *Let f be a function whose domain includes an open interval containing the point x . The derivative of f at x is given by*

$$\frac{d}{dx}f(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{(x+h) - x} = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

There are two common notations for derivatives:

- *Leibniz Notation:* $\frac{d}{dx}(f(x))$
- *Prime or Lagrange Notation:* $f'(x)$



If $f(x)$ is a straight line, the derivative is the slope. For a curve, the slope changes by the values of x , so the derivative is the slope of the line tangent to the curve at x . See, For example, Figure 5.1. If $f'(x)$

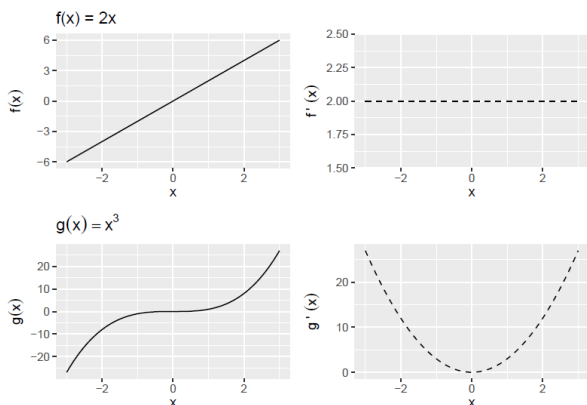


Figure 5.1: The Derivative as a Slope

exists at a point x_0 , then f is said to be **differentiable** at x_0 . That also implies that $f(x)$ is continuous at x_0 .

5.2.1 Properties of derivatives

Suppose that f and g are differentiable at x and that α is a constant. Then the functions $f \pm g$, αf , fg , and f/g (provided $g(x) \neq 0$) are also differentiable at x .

- **Constant rule:**

$$[kf(x)]' = kf'(x)$$

- **Sum rule:**

$$[f(x) \pm g(x)]' = f'(x) \pm g'(x)$$

- **Product rule:**

$$[f(x)g(x)]' = f'(x)g(x) + f(x)g'(x)$$

- **Quotient rule:**

$$[f(x)/g(x)]' = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2}, g(x) \neq 0$$

- **Power rule:**

$$[x^k]' = kx^{k-1}$$

Finally, one way to think of the power of derivatives is that it takes a function a notch down in complexity. The power rule applies to any higher-order function and is proved by induction.

These “rules” become apparent by applying the definition of the derivative above to each of the things to be “derived”, but these come up so frequently that it is best to repeat until it is muscle memory.

Exercise 5.1 (Derivative of Polynomials). For each of the following functions, find the first-order derivative $f'(x)$.

1. $f(x) = c$
2. $f(x) = x$
3. $f(x) = x^2$
4. $f(x) = x^3$
5. $f(x) = \frac{1}{x^2}$
6. $f(x) = (x^3)(2x^4)$
7. $f(x) = x^4 - x^3 + x^2 - x + 1$
8. $f(x) = (x^2 + 1)(x^3 - 1)$
9. $f(x) = 3x^2 + 2x^{1/3}$
10. $f(x) = \frac{x^2+1}{x^2-1}$

◇

5.3 Higher-Order Derivatives (Derivatives of Derivatives)

The first derivative is applying the definition of derivatives on the function, and it can be expressed as

$$f'(x), \quad y', \quad \frac{d}{dx}f(x), \quad \frac{dy}{dx}$$

We can keep applying the differentiation process to functions that are themselves derivatives. The derivative of $f'(x)$ with respect to x , would then be

$$f''(x) = \lim_{h \rightarrow 0} \frac{f'(x+h) - f'(x)}{h}$$

and we can therefore call it the **Second derivative**:

$$f''(x), \quad y'', \quad \frac{d^2}{dx^2}f(x), \quad \frac{d^2y}{dx^2}$$

Similarly, the derivative of $f''(x)$ would be called the third derivative and is denoted $f'''(x)$. And by extension, the **n -th derivative** is expressed as $\frac{d^n}{dx^n}f(x)$, $\frac{d^ny}{dx^n}$.

Example 5.1 (Succession of Derivatives).

$$\begin{aligned}f(x) &= x^3 \\f'(x) &= 3x^2 \\f''(x) &= 6x \\f'''(x) &= 6 \\f''''(x) &= 0\end{aligned}$$

◇

Earlier, in Section 5.2, we said that if a function differentiable at a given point, then it must be continuous. Further, if $f'(x)$ is itself continuous, then $f(x)$ is called continuously differentiable. All of this matters because many of our findings in optimization (Section 6) rely on differentiation, and so we want our function to be differentiable in as many layers. A function that is continuously differentiable infinitely is called “smooth”. Some examples: $f(x) = x^2$, $f(x) = e^x$.

5.4 Composite Functions and the Chain Rule

The rules in the previous section are useful for simple functions, but many functions you’ll see won’t fit neatly in each case immediately. In-

stead, they will be functions of functions, e.g. composite functions. For example, the difference between $x^2 + 1^2$ and $(x^2 + 1)^2$ may look trivial, but the sum rule can be easily applied to the former, while it's actually not obvious what do with the latter.

Composite functions are formed by substituting one function into another and are denoted by

$$(f \circ g)(x) = f[g(x)].$$

To form $f[g(x)]$, the range of g must be contained (at least in part) within the domain of f . The domain of $f \circ g$ consists of all the points in the domain of g for which $g(x)$ is in the domain of f .

Example 5.2. Let $f(x) = \log x$ with range $0 < x < \infty$ and $g(x) = x^2$ with range $-\infty < x < \infty$. Then

$$(f \circ g)(x) = \log x^2, -\infty < x < \infty$$

Also

$$(g \circ f)(x) = [\log x]^2, 0 < x < \infty$$

Notice that $f \circ g$ and $g \circ f$ are not the same functions. ◇

With the notation of composite functions in place, now we can introduce a helpful additional rule that will deal with a derivative of composite functions as a chain of concentric derivatives.

Chain Rule: Let $y = (f \circ g)(x) = f[g(x)]$. The derivative of y with respect to x is

$$\frac{d}{dx}\{f[g(x)]\} = f'[g(x)]g'(x)$$

We can read this as: “the derivative of the composite function y is the derivative of f evaluated at $g(x)$, times the derivative of g .”

The chain rule can be thought of as the derivative of the “outside” times the derivative of the “inside”, remembering that the derivative of the outside function is evaluated at the value of the inside function.

The chain rule can also be written as

$$\frac{dy}{dx} = \frac{dy}{dg(x)} \frac{dg(x)}{dx}$$

This expression does not imply that the $dg(x)$'s cancel out, as in fractions. They are part of the derivative notation and you can't separate them out or cancel them.

Example 5.3 (Composite Exponent). Find $f'(x)$ for $f(x) = (3x^2 + 5x - 7)^6$. ◇

The direct use of a chain rule is when the exponent of is itself a function, so the power rule could not have applied generally.

Generalized Power Rule: If $f(x) = [g(x)]^p$ for any rational number p ,

$$f'(x) = p[g(x)]^{p-1}g'(x)$$

5.5 Derivatives of natural logs and the exponent

Natural logs and exponents (they are inverses of each other; see Section 3.3) crop up everywhere in statistics. Their derivative is a special case from the above, but quite elegant.

Theorem 5.2. *The functions e^x and the natural logarithm $\log(x)$ are continuous and differentiable in their domains, and their first derivate is*

$$(e^x)' = e^x$$

$$\log(x)' = \frac{1}{x}$$

Also, when these are composite functions, it follows by the generalized power rule that

$$\left(e^{g(x)}\right)' = e^{g(x)} \cdot g'(x)$$

$$(\log g(x))' = \frac{g'(x)}{g(x)}, \quad \text{if } g(x) > 0$$



We will relegate the proofs to small excerpts.

5.5.1 Derivatives of natural exponential function

To repeat the main rule in Theorem 5.2, the intuition is that

1. Derivative of e^x is itself: $\frac{d}{dx}e^x = e^x$ (See Figure 5.2)
2. Same thing if there were a constant in front: $\frac{d}{dx}\alpha e^x = \alpha e^x$
3. Same thing no matter how many derivatives there are in front: $\frac{d^n}{dx^n}\alpha e^x = \alpha e^x$
4. Chain Rule: When the exponent is a function of x , remember to take derivative of that function and add to product. $\frac{d}{dx}e^{g(x)} = e^{g(x)}g'(x)$

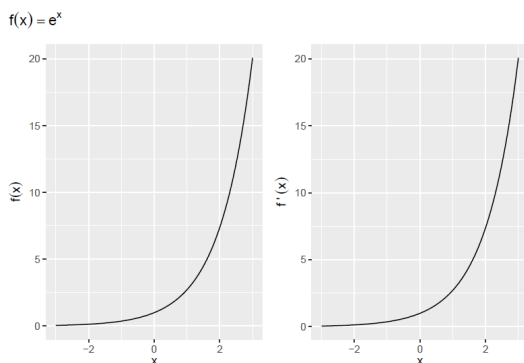


Figure 5.2: Derivative of the Exponential Function

Example 5.4 (Derivative of exponents). Find the derivative for the following.

1. $f(x) = e^{-3x}$
2. $f(x) = e^{x^2}$
3. $f(x) = (x - 1)e^x$

◇

5.5.2 Derivatives of log

The natural log is the mirror image of the natural exponent and has mirroring properties, again, to repeat the theorem,

1. log “prime” x is one over x : $\frac{d}{dx} \log x = \frac{1}{x}$ (Figure 5.3)
2. Exponents become multiplicative constants: $\frac{d}{dx} \log x^k = \frac{d}{dx} k \log x = \frac{k}{x}$
3. Chain rule again: $\frac{d}{dx} \log u(x) = \frac{u'(x)}{u(x)}$
4. For any positive base b , $\frac{d}{dx} b^x = (\log b) (b^x)$.

Example 5.5 (Derivatives of logs). Find dy/dx for the following.

1. $f(x) = \log(x^2 + 9)$
2. $f(x) = \log(\log x)$
3. $f(x) = (\log x)^2$

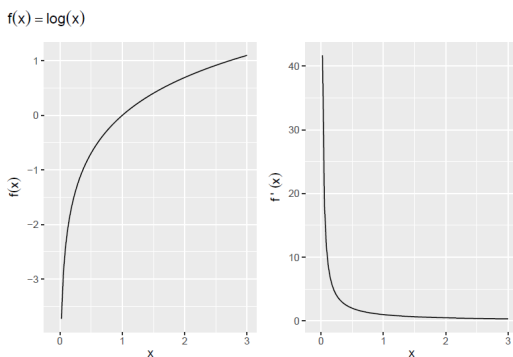


Figure 5.3: Derivative of the Natural Log

4. $f(x) = \log e^x$

◇

5.5.3 Pseudo-Math Proof

We actually show the derivative of the log first, and then the derivative of the exponential naturally follows.

The general derivative of the log at any base a is solvable by the definition of derivatives.

$$(\log_a x)' = \lim_{h \rightarrow 0} \frac{1}{h} \log_a \left(1 + \frac{h}{x} \right)$$

Re-express $g = \frac{h}{x}$ and get

$$\begin{aligned} (\log_a x)' &= \frac{1}{x} \lim_{g \rightarrow 0} \log_a (1 + g)^{\frac{1}{g}} \\ &= \frac{1}{x} \log_a e \end{aligned}$$

By definition of e . As a special case, when $a = e$, then $(\log x)' = \frac{1}{x}$. Now let's think about the inverse, taking the derivative of $y = a^x$.

$$\begin{aligned} y &= a^x \Rightarrow \log y = x \log a \\ &\Rightarrow \frac{y'}{y} = \log a \\ &\Rightarrow y' = y \log a \end{aligned}$$

Then in the special case where $a = e$,

$$(e^x)' = (e^x)$$

5.6 Partial Derivatives

What happens when more than one variable is changing? If you can do ordinary derivatives, you can do partial derivatives: just hold all the other input variables constant except for the one you're differentiating with respect to.

Suppose we have a function $f(x, y)$ now of two (or more) variables and we want to determine the rate of change relative to one of the variables, i.e. derivative of $f(x, y)$ with respect to only y which is denoted as $\frac{\partial f}{\partial y}(x, y)$. To do so, we would find its partial derivative, which is defined similar to the derivative of a function of one variable.

Partial Derivative: Let f be a function of the variables (x_1, \dots, x_n) . The partial derivative of f with respect to x_i is

$$\frac{\partial f}{\partial x_i}(x_1, \dots, x_n) = \lim_{h \rightarrow 0} \frac{f(x_1, \dots, x_i + h, \dots, x_n) - f(x_1, \dots, x_i, \dots, x_n)}{h}$$

Only the i th variable changes — the others are treated as constants.

We can take higher-order partial derivatives, like we did with functions of a single variable, except now the higher-order partials can be with respect to multiple variables.

Example 5.6 (More than one type of partial). Notice that you can take partials with regard to different variables. Suppose $f(x, y) = x^2 + y^2$. Then

$$\begin{aligned}\frac{\partial f}{\partial x}(x, y) &= \\ \frac{\partial f}{\partial y}(x, y) &= \\ \frac{\partial^2 f}{\partial x^2}(x, y) &= \\ \frac{\partial^2 f}{\partial x \partial y}(x, y) &= \end{aligned}$$

◇

Exercise 5.3. Let $f(x, y) = x^3y^4 + e^x - \log y$. What are the following partial derivatives?

$$\begin{aligned}\frac{\partial f}{\partial x}(x, y) &= \\ \frac{\partial f}{\partial y}(x, y) &= \\ \frac{\partial^2 f}{\partial x^2}(x, y) &= \\ \frac{\partial^2 f}{\partial x \partial y}(x, y) &= \end{aligned}$$

5.7 Taylor Series Approximation

A common form of approximation used in statistics involves derivatives. A Taylor series is a way to represent common functions as infinite series (a sum of infinite elements) of the function's derivatives at some point a .

For example, Taylor series are very helpful in representing nonlinear (read: difficult) functions as linear (read: manageable) functions. One can thus **approximate** functions by using lower-order, finite series known as **Taylor polynomials**. If $a = 0$, the series is called a Maclaurin series.

Specifically, a Taylor series of a real or complex function $f(x)$ that is infinitely differentiable in the neighborhood of point a is:

$$\begin{aligned} f(x) &= f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \end{aligned}$$

Taylor Approximation: We can often approximate the curvature of a function $f(x)$ at point a using a 2nd order Taylor polynomial around point a :

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + R_2$$

where R_2 is the remainder (R for remainder, 2 for the fact that we took two derivatives) and often treated as negligible, giving us:

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2$$

The more derivatives that are added, the smaller the remainder R and the more accurate the approximation. Proofs involving limits guarantee that the remainder converges to 0 as the order of derivation increases.

5.8 The Indefinite Integration

So far, we've been interested in finding the derivative $f = F'$ of a function F . However, sometimes we're interested in exactly the reverse: finding the function F for which f is its derivative. We refer to F as the antiderivative of f .

Definition 5.2 (Antiderivative). *The antiderivative of a function $f(x)$ is a differentiable function F whose derivative is f .*

$$F' = f.$$



Another way to describe this is through the inverse formula. Let DF be the derivative of F . And let $DF(x)$ be the derivative of F evaluated at x . Then the antiderivative is denoted by D^{-1} (i.e., the inverse derivative). If $DF = f$, then $F = D^{-1}f$.

This definition bolsters the main takeaway about integrals and derivatives: they are inverses of each other.

Exercise 5.4 (Antiderivative). Find the antiderivative of the following:

1. $f(x) = \frac{1}{x^2}$
2. $f(x) = 3e^{3x}$



We know from derivatives how to manipulate F to get f . But how do you express the procedure to manipulate f to get F ? For that, we need a new symbol, which we will call indefinite integration.

Definition 5.3 (Indefinite Integral). *The indefinite integral of $f(x)$ is written*

$$\int f(x)dx$$

and is equal to the antiderivative of f .



Example 5.7. Draw the function $f(x)$ and its indefinite integral, $\int f(x)dx$

$$f(x) = (x^2 - 4)$$

Solution. The Indefinite Integral of the function $f(x) = (x^2 - 4)$ can, for example, be $F(x) = \frac{1}{3}x^3 - 4x$. But it can also be $F(x) = \frac{1}{3}x^3 - 4x + 1$, because the constant 1 disappears when taking the derivative. ◇

Some of these functions are plotted in the lower panel of Figure 5.4 as dotted lines. Notice from these examples that while there is only a single derivative for any function, there are multiple antiderivatives: one for any arbitrary constant c . c just shifts the curve up or down on the y -axis. If more information is present about the antiderivative — e.g., that it passes through a particular point — then we can solve for a specific value of c .

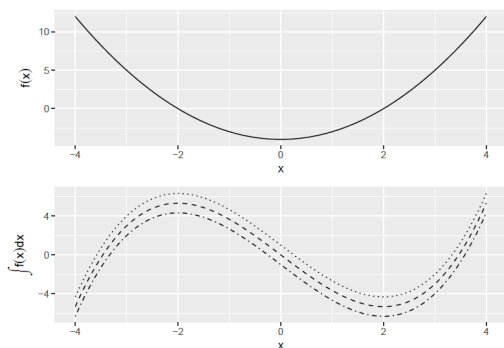


Figure 5.4: The Many Indefinite Integrals of a Function

5.8.1 Common Rules of Integration

Some common rules of integrals follow by virtue of being the inverse of a derivative.

1. Constants are allowed to move out of the integral: $\int af(x)dx = a \int f(x)dx$
2. Integration of the sum, is a sum of integrations: $\int [f(x)+g(x)]dx = \int f(x)dx + \int g(x)dx$
3. Reverse Power-rule: $\int x^n dx = \frac{1}{n+1}x^{n+1} + c$
4. Exponents are still exponents: $\int e^x dx = e^x + c$
5. Recall the derivative of $\log(x)$ is one over x , and so: $\int \frac{1}{x} dx = \log x + c$
6. Reverse chain-rule: $\int e^{f(x)} f'(x) dx = e^{f(x)} + c$
7. More generally: $\int [f(x)]^n f'(x) dx = \frac{1}{n+1} [f(x)]^{n+1} + c$
8. Remember the derivative of a log of a function: $\int \frac{f'(x)}{f(x)} dx = \log f(x) + c$

Example 5.8 (Common Integration). Simplify the following indefinite integrals:

- $\int 3x^2 dx =$
- $\int (2x + 1) dx =$
- $\int e^x e^{e^x} dx =$

◇

5.9 The Definite Integral: The Area under the Curve

If there is a indefinite integral, there *must* be a definite integral. Indeed there is, but the notion of definite integrals comes from a different objective: finding the area under a function. We will find, perhaps remarkably, that the formula to get the sum turns out to be expressible by the anti-derivative.

Suppose we want to determine the area $A(R)$ of a region R defined by a curve $f(x)$ and some interval $a \leq x \leq b$.

One way to calculate the area would be to divide the interval $a \leq x \leq b$ into n subintervals of length Δx and then approximate the region with a series of rectangles, where the base of each rectangle is Δx and the height is $f(x)$ at the midpoint of that interval. $A(R)$ would then be approximated by the area of the union of the rectangles, which is given by

$$S(f, \Delta x) = \sum_{i=1}^n f(x_i) \Delta x$$

and is called a **Riemann sum**.

As we decrease the size of the subintervals Δx , making the rectangles “thinner,” we would expect our approximation of the area of the region to become closer to the true area. This allows us to express the area as a limit of a series:

$$A(R) = \lim_{\Delta x \rightarrow 0} \sum_{i=1}^n f(x_i) \Delta x$$

This is how we define the “Definite” Integral:

Definition 5.4 (The Definite Integral (Riemann)). *If for a given function f the Riemann sum approaches a limit as $\Delta x \rightarrow 0$, then that limit is called the Riemann integral of f from a to b . We express this with the \int , symbol, and write*

$$\int_a^b f(x) dx = \lim_{\Delta x \rightarrow 0} \sum_{i=1}^n f(x_i) \Delta x$$

That is, we read

$$\int_a^b f(x) dx$$

as the definite integral of f from a to b and we defined as the area under the “curve” $f(x)$ from point $x = a$ to $x = b$. ♠

The fundamental theorem of calculus shows us that this sum is, in fact, the antiderivative.

Theorem 5.5 (First Fundamental Theorem of Calculus). *Let the function f be bounded on $[a, b]$ and continuous on (a, b) . Then, suggestively, use the symbol $F(x)$ to denote the definite integral from a to x :*

$$F(x) = \int_a^x f(t)dt, \quad a \leq x \leq b$$

Then $F(x)$ has a derivative at each point in (a, b) and

$$F'(x) = f(x), \quad a < x < b$$

That is, the definite integral function of f is the one of the antiderivatives of some f . ♠

In other words differentiation is the inverse of integration. But now, we've covered definite integrals. The second theorem gives us a simple way of computing a definite integral as a function of indefinite integrals.

Theorem 5.6 (Second Fundamental Theorem of Calculus). *Let the function f be bounded on $[a, b]$ and continuous on (a, b) . Let F be any function that is continuous on $[a, b]$ such that $F'(x) = f(x)$ on (a, b) . Then*

$$\int_a^b f(x)dx = F(b) - F(a)$$

♠

So the procedure to calculate a simple definite integral $\int_a^b f(x)dx$ is then

1. Find the indefinite integral $F(x)$.
2. Evaluate $F(b) - F(a)$.

Example 5.9 (Definite Integral of a monomial). Solve $\int_1^3 3x^2 dx$.

Let $f(x) = 3x^2$.

◇

Exercise 5.7. What is the value of $\int_{-2}^2 e^x e^{e^x} dx$?

◇

5.9.1 Common Rules for Definite Integrals

The area-interpretation of the definite integral provides some rules for simplification.

1. There is no area below a point:

$$\int_a^a f(x)dx = 0$$

2. Reversing the limits changes the sign of the integral:

$$\int_a^b f(x)dx = - \int_b^a f(x)dx$$

3. Sums can be separated into their own integrals:

$$\int_a^b [\alpha f(x) + \beta g(x)]dx = \alpha \int_a^b f(x)dx + \beta \int_a^b g(x)dx$$

4. Areas can be combined as long as limits are linked:

$$\int_a^b f(x)dx + \int_b^c f(x)dx = \int_a^c f(x)dx$$

Exercise 5.8. Simplify the following definite integrals.

1. $\int_1^1 3x^2 dx =$

2. $\int_0^4 (2x + 1)dx =$

3. $\int_{-2}^0 e^x e^{e^x} dx + \int_0^2 e^x e^{e^x} dx =$

◇

5.10 Integration by Substitution

From the second fundamental theorem of calculus, we know that a quick way to get a definite integral is to first find the indefinite integral, and then just plug in the bounds.

Sometimes the integrand (the thing that we are trying to take an integral of) doesn't appear integrable using common rules and antiderivatives. A method one might try is **integration by substitution**, which is related to the Chain Rule.

Suppose we want to find the indefinite integral

$$\int g(x)dx$$

but $g(x)$ is complex and none of the formulas we have seen so far seem to apply immediately. The trick is to come up with a *new* function $u(x)$ such that

$$g(x) = f[u(x)]u'(x).$$

Why does an introduction of yet another function end up simplifying things? Let's refer to the antiderivative of f as F . Then the chain rule tells us that

$$\frac{d}{dx}F[u(x)] = f[u(x)]u'(x).$$

So, $F[u(x)]$ is the antiderivative of g . We can then write

$$\int g(x)dx = \int f[u(x)]u'(x)dx = \int \frac{d}{dx}F[u(x)]dx = F[u(x)] + c$$

To summarize, the procedure to determine the indefinite integral $\int g(x)dx$ by the method of substitution:

1. Identify some part of $g(x)$ that might be simplified by substituting in a single variable u (which will then be a function of x).
2. Determine if $g(x)dx$ can be reformulated in terms of u and du .
3. Solve the indefinite integral.
4. Substitute back in for x

Substitution can also be used to calculate a definite integral. Using the same procedure as above,

$$\int_a^b g(x)dx = \int_c^d f(u)du = F(d) - F(c)$$

where $c = u(a)$ and $d = u(b)$.

Example 5.10 (Integration by Substitution I). Solve the indefinite integral

$$\int x^2 \sqrt{x+1} dx.$$

◇

For the above problem, we could have also used the substitution $u = \sqrt{x+1}$. Then $x = u^2 - 1$ and $dx = 2u du$. Substituting these in, we get

$$\int x^2 \sqrt{x+1} dx = \int (u^2 - 1)^2 u 2u du$$

which when expanded is again a polynomial and gives the same result as above.

Another case in which integration by substitution is useful is with a fraction.

Example 5.11 (Integration by Substitution II). Simplify

$$\int_0^1 \frac{5e^{2x}}{(1+e^{2x})^{1/3}} dx.$$

◇

5.11 Integration by Parts

Another useful integration technique is **integration by parts**, which is related to the Product Rule of differentiation. The product rule states that

$$\frac{d}{dx}(uv) = u \frac{dv}{dx} + v \frac{du}{dx}$$

Integrating this and rearranging, we get

$$\int u \frac{dv}{dx} dx = uv - \int v \frac{du}{dx} dx$$

or

$$\int u(x)v'(x)dx = u(x)v(x) - \int v(x)u'(x)dx$$

More easily remembered with the mnemonic “Ultraviolet Voodoo”:

$$\int u dv = uv - \int v du$$

where $du = u'(x)dx$ and $dv = v'(x)dx$.

For definite integrals, this is simply

$$\int_a^b u \frac{dv}{dx} dx = uv \Big|_a^b - \int_a^b v \frac{du}{dx} dx$$

Our goal here is to find expressions for u and dv that, when substituted into the above equation, yield an expression that's more easily evaluated.

Example 5.12 (Integration by Parts). Simplify the following integrals. These seemingly obscure forms of integrals come up often when integrating distributions.

$$\int x e^{ax} dx$$

Solution. Let $u = x$ and $\frac{dv}{dx} = e^{ax}$. Then $du = dx$ and $v = (1/a)e^{ax}$. Substituting this into the integration by parts formula, we obtain

$$\begin{aligned} \int x e^{ax} dx &= uv - \int v du \\ &= x \left(\frac{1}{a} e^{ax} \right) - \int \frac{1}{a} e^{ax} dx \\ &= \frac{1}{a} x e^{ax} - \frac{1}{a^2} e^{ax} + c \end{aligned}$$

◇

Exercise 5.9 (Integration by Parts II). Integrate:

1. $\int x^n e^{ax} dx$
2. $\int x^3 e^{-x^2} dx$

◇

Chapter 6

Optimization

To optimize, we use derivatives and calculus. Optimization is to find the maximum or minimum of a function, and to find what value of an input gives that extrema. This has obvious uses in engineering. Many tools in the statistical toolkit use optimization. One of the most common ways of estimating a model is through “Maximum Likelihood Estimation”, done via optimizing a function (the likelihood).

Optimization is common in Machine Learning, Satellite Navigation (SatNav) systems, Business Management, etc.. A go-to model of human behavior is that they optimize a certain utility function – for example, why use a walk on a concrete walkway with a 90 degree path, when you can simply walk the diagonal across a patch of grass. Humans are not pure utility maximizers, of course, but nuanced models of optimization – for example, adding constraints and adding uncertainty – will prove to be quite useful.

6.1 Maxima and Minima

The first derivative, $f'(x)$, quantifies the slope of a function. Therefore, it can be used to check whether the function $f(x)$ at the point x is increasing or decreasing at x .

1. **Increasing:** $f'(x) > 0$
2. **Decreasing:** $f'(x) < 0$
3. **Neither increasing nor decreasing:** $f'(x) = 0$, i.e. a maximum, minimum, or saddle point

So for example, $f(x) = x^2 + 2$ and $f'(x) = 2x$ shown in Figure 6.1.

The second derivative $f''(x)$ identifies whether the function $f(x)$ at the point x is

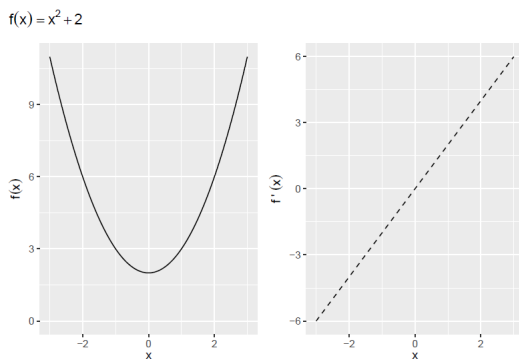


Figure 6.1: Maxima and Minima

1. Concave down: $f''(x) < 0$
2. Concave up: $f''(x) > 0$

Maximum (Minimum): x_0 is a **local maximum (minimum)** if $f(x_0) > f(x)$ ($f(x_0) < f(x)$) for all x within some open interval containing x_0 . x_0 is a **global maximum (minimum)** if $f(x_0) > f(x)$ ($f(x_0) < f(x)$) for all x in the domain of f .

Given the function f defined over domain D , all of the following are defined as **critical points**:

1. Any interior point of D where $f'(x) = 0$.
2. Any interior point of D where $f'(x)$ does not exist.
3. Any endpoint that is in D .

The maxima and minima will be a subset of the critical points.

Second Derivative Test of Maxima/Minima: We can use the second derivative to tell us whether a point is a maximum or minimum of $f(x)$.

1. Local Maximum: $f'(x) = 0$ and $f''(x) < 0$
2. Local Minimum: $f'(x) = 0$ and $f''(x) > 0$
3. Need more info: $f'(x) = 0$ and $f''(x) = 0$

Global Maxima and Minima Sometimes no global max or min exists — e.g., $f(x)$ not bounded above or below. However, there are three situations where we can fairly easily identify global max or min.

1. **Functions with only one critical point.** If x_0 is a local max or min of f and it is the only critical point, then it is the global max or min.
2. **Globally concave up or concave down functions.** If $f''(x)$ is never zero, then there is at most one critical point. That critical point is a global maximum if $f'' < 0$ and a global minimum if $f'' > 0$.
3. **Functions over closed and bounded intervals** must have both a global maximum and a global minimum.

6.2 Concavity of a Function

Concavity helps identify the curvature of a function, $f(x)$, in 2-dimensional space.

Definition 6.1 (Concave Function). A function f is strictly concave over the set S if $\forall x_1, x_2 \in S$ and $\forall a \in (0, 1)$,

$$f(ax_1 + (1 - a)x_2) > af(x_1) + (1 - a)f(x_2)$$

Any line connecting two points on a concave function will lie below the function. See the left panel of Figure 6.2. ♠

Definition 6.2 (Convex Function). Convex: A function f is strictly convex over the set S if $\forall x_1, x_2 \in S$ and $\forall a \in (0, 1)$,

$$f(ax_1 + (1 - a)x_2) < af(x_1) + (1 - a)f(x_2)$$

Any line connecting two points on a convex function will lie above the function. See the right panel of Figure 6.2. ♠

Sometimes, concavity and convexity are strict of a requirement. For most purposes of getting solutions, what we call quasi-concavity is enough.

Definition 6.3 (Quasiconcave Function). A function f is quasiconcave over the set S if $\forall x_1, x_2 \in S$ and $\forall a \in (0, 1)$,

$$f(ax_1 + (1 - a)x_2) \geq \min(f(x_1), f(x_2))$$

No matter what two points you select, the lowest valued point will always be an end point. ♠

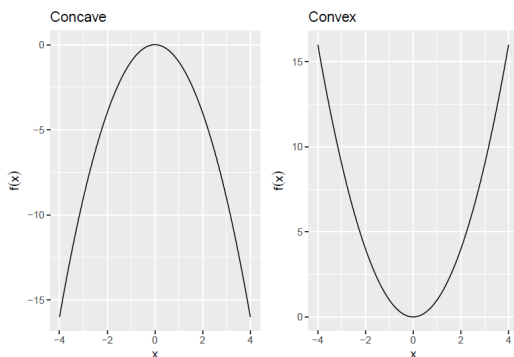


Figure 6.2: Concave and Convex functions $f(x)$

Definition 6.4 (Quasiconvex Function). A function f is quasiconvex over the set S if $\forall x_1, x_2 \in S$ and $\forall a \in (0, 1)$,

$$f(ax_1 + (1 - a)x_2) \leq \max(f(x_1), f(x_2))$$

No matter what two points you select, the highest valued point will always be an end point.



Second Derivative Test of Concavity: The second derivative can be used to understand concavity. If

$$\begin{aligned} f''(x) < 0 &\Rightarrow \text{Concave} \\ f''(x) > 0 &\Rightarrow \text{Convex} \end{aligned}$$

6.2.1 Quadratic Forms

Quadratic forms is shorthand for a way to summarize a function. This is important for finding concavity because

1. Approximates local curvature around a point — e.g., used to identify max vs min vs saddle point.
2. They are simple to express even in n dimensions:
3. Have a matrix representation.

Quadratic Form: A polynomial where each term is a monomial of degree 2 in any number of variables:

$$\text{One variable: } Q(x_1) = a_{11}x_1^2$$

$$\text{Two variables: } Q(x_1, x_2) = a_{11}x_1^2 + a_{12}x_1x_2 + a_{22}x_2^2$$

$$\text{N variables: } Q(x_1, \dots, x_n) = \sum_{i \leq j} a_{ij}x_i x_j$$

which can be written in matrix terms:

One variable

$$Q(\mathbf{x}) = x_1^\top a_{11} x_1$$

N variables:

$$\begin{aligned} Q(\mathbf{x}) &= \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} a_{11} & \frac{1}{2}a_{12} & \cdots & \frac{1}{2}a_{1n} \\ \frac{1}{2}a_{12} & a_{22} & \cdots & \frac{1}{2}a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}a_{1n} & \frac{1}{2}a_{2n} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\ &= \mathbf{x}^\top \mathbf{A} \mathbf{x} \end{aligned}$$

For example, the Quadratic on \mathbf{R}^2 :

$$\begin{aligned} Q(x_1, x_2) &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a_{11} & \frac{1}{2}a_{12} \\ \frac{1}{2}a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= a_{11}x_1^2 + a_{12}x_1x_2 + a_{22}x_2^2 \end{aligned}$$

6.2.2 Definiteness of Quadratic Forms

When the function $f(\mathbf{x})$ has more than two inputs, determining whether it has a maxima and minima (remember, functions may have many inputs but they have only one output) is a bit more tedious. Definiteness helps identify the curvature of a function, $Q(\mathbf{x})$, in n dimensional space.

Definiteness: By definition, a quadratic form always takes on the value of zero when $x = 0$, $Q(\mathbf{x}) = 0$ at $\mathbf{x} = 0$. The definiteness of the matrix \mathbf{A} is determined by whether the quadratic form $Q(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}$ is greater than zero, less than zero, or sometimes both over all $\mathbf{x} \neq 0$.

6.3 FOC and SOC

We can see from a graphical representation that if a point is a local maxima or minima, it must meet certain conditions regarding its derivative. These are so commonly used that we refer these to “First Order Conditions” (FOCs) and “Second Order Conditions” (SOCs).

6.3.1 First Order Conditions

When we examined functions of one variable x , we found critical points by taking the first derivative, setting it to zero, and solving for x . For functions of n variables, the critical points are found in much the same way, except now we set the partial derivatives equal to zero. Note: We will only consider critical points on the interior of a function's domain.

In a derivative, we only took the derivative with respect to one variable at a time. When we take the derivative separately with respect to all variables in the elements of \mathbf{x} and then express the result as a vector, we use the term Gradient and Hessian.

Definition 6.5 (Gradient). *Given a function $f(\mathbf{x})$ in n variables, the gradient $\nabla f(\mathbf{x})$ (the greek letter nabla) is a column vector, where the i th element is the partial derivative of $f(\mathbf{x})$ with respect to x_i :*

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{pmatrix}$$



Before we know whether a point is a maxima or minima, if it meets the FOC it is a “Critical Point”.

Definition 6.6 (Critical Point). *\mathbf{x}^* is a critical point if and only if $\nabla f(\mathbf{x}^*) = 0$. If the partial derivative of $f(x)$ with respect to x^* is 0, then \mathbf{x}^* is a critical point. To solve for \mathbf{x}^* , find the gradient, set each element equal to 0, and solve the system of equations.*

$$\mathbf{x}^* = \begin{pmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_n^* \end{pmatrix}$$



Example 6.1. Example: Given a function $f(\mathbf{x}) = (x_1 - 1)^2 + x_2^2 + 1$, find the (1) Gradient and (2) Critical point of $f(\mathbf{x})$.

Solution. Gradient

$$\begin{aligned}\nabla f(\mathbf{x}) &= \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \end{pmatrix} \\ &= \begin{pmatrix} 2(x_1 - 1) \\ 2x_2 \end{pmatrix}\end{aligned}$$

Critical Point $\mathbf{x}^* =$

$$\begin{aligned}\frac{\partial f(\mathbf{x})}{\partial x_1} &= 2(x_1 - 1) = 0 \\ \Rightarrow x_1^* &= 1 \\ \frac{\partial f(\mathbf{x})}{\partial x_2} &= 2x_2 = 0 \\ \Rightarrow x_2^* &= 0\end{aligned}$$

So

$$\mathbf{x}^* = (1, 0)$$

◇

6.3.2 Second Order Conditions

When we found a critical point for a function of one variable, we used the second derivative as a indicator of the curvature at the point in order to determine whether the point was a min, max, or saddle (second derivative test of concavity). For functions of n variables, we use *second order partial derivatives* as an indicator of curvature.

Definition 6.7 (Hessian). *Given a function $f(\mathbf{x})$ in n variables, the Hessian $\mathbf{H}(\mathbf{x})$ is an $n \times n$ matrix, where the (i, j) th element is the second order partial derivative of $f(\mathbf{x})$ with respect to x_i and x_j :*

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{pmatrix}$$

♠

Note that the Hessian will be a symmetric matrix because $\frac{\partial f(\mathbf{x})}{\partial x_1 \partial x_2} = \frac{\partial f(\mathbf{x})}{\partial x_2 \partial x_1}$. Also note that given that $f(\mathbf{x})$ is of quadratic form, each element

of the Hessian will be a constant. These definitions will be employed when we determine the **Second Order Conditions** of a function:

Given a function $f(\mathbf{x})$ and a point \mathbf{x}^* such that $\nabla f(\mathbf{x}^*) = 0$,

1. Hessian is Positive Definite \implies Strict Local Min
2. Hessian is Positive Semidefinite $\forall \mathbf{x} \in B(\mathbf{x}^*, \epsilon)\} \implies$ Local Min
3. Hessian is Negative Definite \implies Strict Local Max
4. Hessian is Negative Semidefinite $\forall \mathbf{x} \in B(\mathbf{x}^*, \epsilon)\} \implies$ Local Max
5. Hessian is Indefinite \implies Saddle Point

Example 6.2 (Max, min, and definiteness). We found that the only critical point of $f(\mathbf{x}) = (x_1 - 1)^2 + x_2^2 + 1$ is at $\mathbf{x}^* = (1, 0)$. Is it a min, max, or saddle point?

Solution. The Hessian is

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

The Leading principal minors of the Hessian are $M_1 = 2; M_2 = 4$. Now we consider Definiteness. Since both leading principal minors are positive, the Hessian is positive definite.

Maxima, Minima, or Saddle Point? Since the Hessian is positive definite and the gradient equals 0, $\mathbf{x}^* = (1, 0)$ is a strict local minimum.

Note: Alternate check of definiteness. Is $\mathbf{H}(\mathbf{x}^*) \succeq 0 \quad \forall \quad \mathbf{x} \neq 0$

$$\begin{aligned} \mathbf{x}^\top \mathbf{H}(\mathbf{x}^*) \mathbf{x} &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \\ &\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 2x_1^2 + 2x_2^2 \end{aligned}$$

For any $\mathbf{x} \neq 0$, $2(x_1^2 + x_2^2) > 0$, so the Hessian is positive definite and \mathbf{x}^* is a strict local minimum. \diamond

6.3.3 Definiteness and Concavity

Although definiteness helps us to understand the curvature of an n -dimensional function, it does not necessarily tell us whether the function is globally concave or convex.

We need to know whether a function is globally concave or convex to determine whether a critical point is a global min or max. We can use the definiteness of the Hessian to determine whether a function is globally concave or convex:

1. Hessian is Positive Semidefinite $\forall \mathbf{x} \implies$ Globally Convex
2. Hessian is Negative Semidefinite $\forall \mathbf{x} \implies$ Globally Concave

Notice that the definiteness conditions must be satisfied over the entire domain.

6.4 Global Maxima and Minima

Global Max/Min Conditions: Given a function $f(\mathbf{x})$ and a point \mathbf{x}^* such that $\nabla f(\mathbf{x}^*) = 0$,

1. $f(\mathbf{x})$ Globally Convex \implies Global Min
2. $f(\mathbf{x})$ Globally Concave \implies Global Max

Note that showing that $\mathbf{H}(\mathbf{x}^*)$ is negative semidefinite is not enough to guarantee \mathbf{x}^* is a local max. However, showing that $\mathbf{H}(\mathbf{x})$ is negative semidefinite for all \mathbf{x} guarantees that \mathbf{x}^* is a global max. The same goes for positive semidefinite and minima.

Example 6.3. Given $f_1(x) = x^4$ and $f_2(x) = -x^4$.

- Both functions have $x = 0$ as a critical point.
- Unfortunately, $f_1''(0) = 0$ and $f_2''(0) = 0$, so we can't tell whether $x = 0$ is a min or max for either.
- However, $f_1''(x) = 12x^2$ and $f_2''(x) = -12x^2$.
- For all x , $f_1''(x) \geq 0$ and $f_2''(x) \leq 0$ — i.e., $f_1(x)$ is globally convex and $f_2(x)$ is globally concave.

So $x = 0$ is a global min of $f_1(x)$ and a global max of $f_2(x)$.

◇

Example 6.4. Given $f(\mathbf{x}) = x_1^3 - x_2^3 + 9x_1x_2$, find any maxima or minima.

Solution. We can use the following steps to help us solve this problem.

1. First order conditions.

(a) Gradient $\nabla f(\mathbf{x}) =$

$$\begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 3x_1^2 + 9x_2 \\ -3x_2^2 + 9x_1 \end{pmatrix}$$

(b) Critical Points $\mathbf{x}^* =$

Set the gradient equal to zero and solve for x_1 and x_2 . We have two equations and two unknowns. Solving for x_1 and x_2 , we get two critical points: $\mathbf{x}_1^* = (0, 0)$ and $\mathbf{x}_2^* = (3, -3)$.

$$\begin{aligned} 3x_1^2 + 9x_2 = 0 &\Rightarrow 9x_2 = -3x_1^2 \\ &\Rightarrow x_2 = -\frac{1}{3}x_1^2 \\ -3x_2^2 + 9x_1 = 0 &\Rightarrow -3\left(-\frac{1}{3}x_1^2\right)^2 + 9x_1 = 0 \\ &\Rightarrow -\frac{1}{3}x_1^4 + 9x_1 = 0 \\ &\Rightarrow x_1^3 = 27x_1 \\ &\Rightarrow x_1 = 3 \\ 3(3)^2 + 9x_2 = 0 &\Rightarrow x_2 = -3 \end{aligned}$$

2. Second order conditions.

(a) Hessian $\mathbf{H}(\mathbf{x}) =$

$$\begin{pmatrix} 6x_1 & 9 \\ 9 & -6x_2 \end{pmatrix}$$

(b) Hessian $\mathbf{H}(\mathbf{x}_1^*) =$

$$\begin{pmatrix} 0 & 9 \\ 9 & 0 \end{pmatrix}$$

(c) Leading principal minors of $\mathbf{H}(\mathbf{x}_1^*) =$

$$M_1 = 0; M_2 = -81$$

(d) Definiteness of $\mathbf{H}(\mathbf{x}_1^*)?$

$\mathbf{H}(\mathbf{x}_1^*)$ is indefinite

(e) Maxima, Minima, or Saddle Point for \mathbf{x}_1^* ?

Since $\mathbf{H}(\mathbf{x}_1^*)$ is indefinite, $\mathbf{x}_1^* = (0, 0)$ is a saddle point.

(f) Hessian $\mathbf{H}(\mathbf{x}_2^*) =$

$$\begin{pmatrix} 18 & 9 \\ 9 & 18 \end{pmatrix}$$

(g) Leading principal minors of $\mathbf{H}(\mathbf{x}_2^*) =$

$$M_1 = 18; M_2 = 243$$

(h) Definiteness of $\mathbf{H}(\mathbf{x}_2^*)?$

$\mathbf{H}(\mathbf{x}_2^*)$ is positive definite

- (i) Maxima, Minima, or Saddle Point for \mathbf{x}_2^* ?
 Since $\mathbf{H}(\mathbf{x}_2^*)$ is positive definite, $\mathbf{x}_1^* = (3, -3)$ is a strict local minimum

3. Global concavity/convexity.

- (a) Is $f(\mathbf{x})$ globally concave/convex?
 No. In evaluating the Hessians for \mathbf{x}_1^* and \mathbf{x}_2^* we saw that the Hessian is not positive semidefinite at $\mathbf{x} = (0,0)$.
- (b) Are any \mathbf{x}^* global minima or maxima?
 No. Since the function is not globally concave/convex, we can't infer that $\mathbf{x}_2^* = (3, -3)$ is a global minimum. In fact, if we set $x_1 = 0$, the $f(\mathbf{x}) = -x_2^3$, which will go to $-\infty$ as $x_2 \rightarrow \infty$.

◇

6.5 Constrained Optimization

We have already looked at optimizing a function in one or more dimensions over the whole domain of the function. Often, however, we want to find the maximum or minimum of a function over some restricted part of its domain.

Types of Constraints: For a function $f(x_1, \dots, x_n)$, there are two types of constraints that can be imposed:

1. **Equality constraints:** constraints of the form $c(x_1, \dots, x_n) = r$. Budget constraints are the classic example of equality constraints in data science.
2. **Inequality constraints:** constraints of the form $c(x_1, \dots, x_n) \leq r$. These might arise from non-negativity constraints or other threshold effects.

In any constrained optimization problem, the constrained maximum will always be less than or equal to the unconstrained maximum. If the constrained maximum is less than the unconstrained maximum, then the constraint is binding. Essentially, this means that you can treat your constraint as an equality constraint rather than an inequality constraint.

For example, the budget constraint binds when you spend your entire budget. This generally happens because we believe that utility is strictly increasing in consumption, i.e. you always want more so you spend everything you have.

Any number of constraints can be placed on an optimization problem. When working with multiple constraints, always make sure that the set

of constraints are not pathological; it must be possible for all of the constraints to be satisfied simultaneously.

Set-up for Constrained Optimization:

$$\max_{x_1, x_2} f(x_1, x_2) \text{ s.t. } c(x_1, x_2)$$

$$\min_{x_1, x_2} f(x_1, x_2) \text{ s.t. } c(x_1, x_2)$$

This tells us to maximize/minimize our function, $f(x_1, x_2)$, with respect to the choice variables, x_1, x_2 , subject to the constraint.

Example 6.5.

$$\max_{x_1, x_2} f(x_1, x_2) = -(x_1^2 + 2x_2^2) \text{ s.t. } x_1 + x_2 = 4$$

It is easy to see that the *unconstrained* maximum occurs at $(x_1, x_2) = (0, 0)$, but that does not satisfy the constraint. How should we proceed?

◇

6.5.1 Equality Constraints

Equality constraints are the easiest to deal with because we know that the maximum or minimum has to lie on the (intersection of the) constraint(s).

The trick is to change the problem from a constrained optimization problem in n variables to an unconstrained optimization problem in $n+k$ variables, adding *one* variable for *each* equality constraint. We do this using a Lagrangian multiplier.

Lagrangian function: The Lagrangian function allows us to combine the function we want to optimize and the constraint function into a single function. Once we have this single function, we can proceed as if this were an *unconstrained* optimization problem.

For each constraint, we must include a **Lagrange multiplier** (λ_i) as an additional variable in the analysis. These terms are the link between the constraint and the Lagrangian function.

Given a *two dimensional* set-up:

$$\max_{x_1, x_2} / \min_{x_1, x_2} f(x_1, x_2) \text{ s.t. } c(x_1, x_2) = a$$

We define the Lagrangian function $L(x_1, x_2, \lambda_1)$ as follows:

$$L(x_1, x_2, \lambda_1) = f(x_1, x_2) - \lambda_1(c(x_1, x_2) - a)$$

More generally, in n dimensions:

$$L(x_1, \dots, x_n, \lambda_1, \dots, \lambda_k) = f(x_1, \dots, x_n) - \sum_{i=1}^k \lambda_i(c_i(x_1, \dots, x_n) - r_i)$$

Getting the sign right: Note that above we subtract the Lagrangian term *and* we subtract the constraint constant from the constraint function. Occasionally, you may see the following alternative form of the Lagrangian, which is *equivalent*:

$$L(x_1, \dots, x_n, \lambda_1, \dots, \lambda_k) = f(x_1, \dots, x_n) + \sum_{i=1}^k \lambda_i (r_i - c_i(x_1, \dots, x_n))$$

Here we add the Lagrangian term *and* we subtract the constraining function from the constraint constant.

Using the Lagrangian to Find the Critical Points: To find the critical points, we take the partial derivatives of lagrangian function, $L(x_1, \dots, x_n, \lambda_1, \dots, \lambda_k)$, with respect to each of its variables (all choice variables \mathbf{x} *and* all Lagrangian multipliers λ). At a critical point, each of these partial derivatives must be equal to zero, so we obtain a system of $n + k$ equations in $n + k$ unknowns:

$$\begin{aligned} \frac{\partial L}{\partial x_1} &= \frac{\partial f}{\partial x_1} - \sum_{i=1}^k \lambda_i \frac{\partial c_i}{\partial x_1} = 0 \\ &\vdots \\ \frac{\partial L}{\partial x_n} &= \frac{\partial f}{\partial x_n} - \sum_{i=1}^k \lambda_i \frac{\partial c_i}{\partial x_n} = 0 \\ \frac{\partial L}{\partial \lambda_1} &= c_1(x_1, \dots, x_n) - r_1 = 0 \\ &\vdots \\ \frac{\partial L}{\partial \lambda_k} &= c_k(x_1, \dots, x_n) - r_k = 0 \end{aligned}$$

We can then solve this system of equations, because there are $n + k$ equations and $n + k$ unknowns, to calculate the critical point $(x_1^*, \dots, x_n^*, \lambda_1^*, \dots, \lambda_k^*)$.

Second-order Conditions and Unconstrained Optimization:

There may be more than one critical point, i.e. we need to verify that the critical point we find is a maximum/minimum. Similar to unconstrained optimization, we can do this by checking the second-order conditions.

Example 6.6 (Constrained optimization with two consumer-products and a budget constraint). Find the constrained optimization of

$$\max_{x_1, x_2} f(x) = -(x_1^2 + 2x_2^2) \text{ s.t. } x_1 + x_2 = 4$$

Solution. Solve by the following steps.

1. Begin by writing the Lagrangian:

$$L(x_1, x_2, \lambda) = -(x_1^2 + 2x_2^2) - \lambda(x_1 + x_2 - 4)$$

2. Take the partial derivatives and set equal to zero:

$$\frac{\partial L}{\partial x_1} = -2x_1 - \lambda = 0$$

$$\frac{\partial L}{\partial x_2} = -4x_2 - \lambda = 0$$

$$\frac{\partial L}{\partial \lambda} = -(x_1 + x_2 - 4) = 0$$

3. Solve the system of equations: Using the first two partials, we see that $\lambda = -2x_1$ and $\lambda = -4x_2$. Set these equal to see that $x_1 = 2x_2$. Using the third partial and the above equality, $4 = 2x_2 + x_2$ from which we get

$$x_2^* = 4/3, \quad x_1^* = 8/3, \quad \lambda = -16/3$$

4. Therefore, the only critical point is $x_1^* = \frac{8}{3}$ and $x_2^* = \frac{4}{3}$. This gives $f(\frac{8}{3}, \frac{4}{3}) = -\frac{96}{9}$, which is less than the unconstrained optimum $f(0, 0) = 0$. \diamond

Notice that when we take the partial derivative of L with respect to the Lagrangian multiplier and set it equal to 0, we return exactly our constraint! This is why signs matter.

6.6 Inequality Constraints

Inequality constraints define the boundary of a region over which we seek to optimize the function. This makes inequality constraints more challenging because we do not know if the maximum/minimum lies along one of the constraints (the constraint binds) or in the interior of the region.

We must introduce more variables in order to turn the problem into an unconstrained optimization.

Slack: For each inequality constraint $c_i(x_1, \dots, x_n) \leq a_i$, we define a slack variable s_i^2 for which the expression $c_i(x_1, \dots, x_n) \leq a_i - s_i^2$ would hold with equality. These slack variables capture how close the constraint comes to binding. We use s^2 rather than s to ensure that the slack is positive. Note: Slack is just a way to transform our constraints.

Given a two-dimensional set-up and these edited constraints:

$$\max_{x_1, x_2} / \min_{x_1, x_2} f(x_1, x_2) \text{ s.t. } c(x_1, x_2) \leq a_1$$

Adding in Slack:

$$\max_{x_1, x_2} / \min_{x_1, x_2} f(x_1, x_2) \text{ s.t. } c(x_1, x_2) \leq a_1 - s_1^2$$

We define the Lagrangian function $L(x_1, x_2, \lambda_1, s_1)$ as follows:

$$L(x_1, x_2, \lambda_1, s_1) = f(x_1, x_2) - \lambda_1(c(x_1, x_2) + s_1^2 - a_1)$$

More generally, in n dimensions:

$$\begin{aligned} L(x_1, \dots, x_n, \lambda_1, \dots, \lambda_k, s_1, \dots, s_k) \\ = f(x_1, \dots, x_n) - \sum_{i=1}^k \lambda_i(c_i(x_1, \dots, x_n) + s_i^2 - a_i) \end{aligned}$$

Finding the Critical Points: To find the critical points, we take the partial derivatives of the Lagrangian function,

$$L(x_1, \dots, x_n, \lambda_1, \dots, \lambda_k, s_1, \dots, s_k),$$

with respect to each of its variables (all choice variables x , all Lagrangian multipliers λ , and all slack variables s). At a critical point, *each* of these partial derivatives must be equal to zero, so we obtain a system of $n + 2k$ equations in $n + 2k$ unknowns:

$$\begin{aligned} \frac{\partial L}{\partial x_1} &= \frac{\partial f}{\partial x_1} - \sum_{i=1}^k \lambda_i \frac{\partial c_i}{\partial x_1} = 0 \\ &\vdots \\ \frac{\partial L}{\partial x_n} &= \frac{\partial f}{\partial x_n} - \sum_{i=1}^k \lambda_i \frac{\partial c_i}{\partial x_n} = 0 \\ \frac{\partial L}{\partial \lambda_1} &= c_1(x_1, \dots, x_n) + s_1^2 - b_1 = 0 \\ &\vdots \\ \frac{\partial L}{\partial \lambda_k} &= c_k(x_1, \dots, x_n) + s_k^2 - b_k = 0 \\ \frac{\partial L}{\partial s_1} &= 2s_1\lambda_1 = 0 \\ &\vdots \\ \frac{\partial L}{\partial s_k} &= 2s_k\lambda_k = 0 \end{aligned}$$

Complementary slackness conditions: The last set of first order conditions of the form $2s_i\lambda_i = 0$ (the partials taken with respect to the slack variables) are known as complementary slackness conditions. These conditions can be satisfied one of three ways:

1. $\lambda_i = 0$ and $s_i \neq 0$: This implies that the slack is positive and thus *the constraint does not bind*.
2. $\lambda_i \neq 0$ and $s_i = 0$: This implies that there is no slack in the constraint and *the constraint does bind*.
3. $\lambda_i = 0$ and $s_i = 0$: In this case, there is no slack but the *constraint binds trivially*, without changing the optimum.

Example 6.7. Find the critical points for the following constrained optimization:

$$\max_{x_1, x_2} f(x) = -(x_1^2 + 2x_2^2) \text{ s.t. } x_1 + x_2 \leq 4$$

Solution. Solve using the following steps:

1. Rewrite with the slack variables:

$$\max_{x_1, x_2} f(x) = -(x_1^2 + 2x_2^2) \text{ s.t. } x_1 + x_2 \leq 4 - s_1^2$$

2. Write the Lagrangian:

$$L(x_1, x_2, \lambda_1, s_1) = -(x_1^2 + 2x_2^2) - \lambda_1(x_1 + x_2 + s_1^2 - 4)$$

3. Take the partial derivatives and set equal to 0:

$$\begin{aligned} \frac{\partial L}{\partial x_1} &= -2x_1 - \lambda_1 = 0 \\ \frac{\partial L}{\partial x_2} &= -4x_2 - \lambda_1 = 0 \\ \frac{\partial L}{\partial \lambda_1} &= -(x_1 + x_2 + s_1^2 - 4) = 0 \\ \frac{\partial L}{\partial s_1} &= -2s_1\lambda_1 = 0 \end{aligned}$$

4. Consider all ways that the complementary slackness conditions are solved:

Hypothesis	s_1	λ_1	x_1	x_2	$f(x_1, x_2)$
$s_1 = 0 \ \lambda_1 = 0$	No solution				
$s_1 \neq 0 \ \lambda_1 = 0$	2	0	0	0	0
$s_1 = 0 \ \lambda_1 \neq 0$	0	$-\frac{16}{3}$	$\frac{8}{3}$	$\frac{4}{3}$	$-\frac{32}{3}$
$s_1 \neq 0 \ \lambda_1 \neq 0$	No solution				

This shows that there are two critical points: $(0, 0)$ and $(\frac{8}{3}, \frac{4}{3})$.

5. Find maximum: Looking at the values of $f(x_1, x_2)$ at the critical points, we see that $f(x_1, x_2)$ is maximized at $x_1^* = 0$ and $x_2^* = 0$.

◇

Exercise 6.1. Find the critical points for the following constrained optimization:

$$\max_{x_1, x_2} f(x) = -(x_1^2 + 2x_2^2) \text{ s.t. } \begin{aligned} x_1 + x_2 &\leq 4 \\ x_1 &\geq 0 \\ x_2 &\geq 0 \end{aligned}$$

6.7 Kuhn-Tucker Conditions

As you can see, this can be a pain. When dealing explicitly with *non-negativity constraints*, this process is simplified by using the Kuhn-Tucker method.

Because the problem of maximizing a function subject to inequality and non-negativity constraints arises frequently in data science, the **Kuhn-Tucker conditions** provides a method that often makes it easier to both calculate the critical points and identify points that are (local) maxima.

Given a *two-dimensional set-up*:

$$\max_{x_1, x_2} / \min_{x_1, x_2} f(x_1, x_2) \text{ s.t. } \begin{aligned} c(x_1, x_2) &\leq a_1 \\ x_1 &\geq 0 \\ gx_2 &\geq 0 \end{aligned}$$

We define the Lagrangian function $L(x_1, x_2, \lambda_1)$ the same as if we did not have the non-negativity constraints:

$$L(x_1, x_2, \lambda_2) = f(x_1, x_2) - \lambda_1(c(x_1, x_2) - a_1)$$

More generally, in n dimensions:

$$L(x_1, \dots, x_n, \lambda_1, \dots, \lambda_k) = f(x_1, \dots, x_n) - \sum_{i=1}^k \lambda_i(c_i(x_1, \dots, x_n) - a_i)$$

Kuhn-Tucker and Complementary Slackness Conditions:

To find the critical points, we first calculate the Kuhn-Tucker conditions by taking the partial derivatives of the Lagrangian function,

$L(x_1, \dots, x_n, \lambda_1, \dots, \lambda_k)$, with respect to each of its variables (all choice variable x and all lagrangian multipliers λ) and we calculate the *complementary slackness conditions* by multiplying each partial derivative by its respective variable *and* include non-negativity conditions for all variables (choice variables x and Lagrangian multipliers λ).

Kuhn-Tucker Conditions

$$\begin{aligned}\frac{\partial L}{\partial x_1} &\leq 0, \dots, \frac{\partial L}{\partial x_n} \leq 0 \\ \frac{\partial L}{\partial \lambda_1} &\geq 0, \dots, \frac{\partial L}{\partial \lambda_m} \geq 0\end{aligned}$$

Complementary Slackness Conditions

$$\begin{aligned}x_1 \frac{\partial L}{\partial x_1} &= 0, \dots, x_n \frac{\partial L}{\partial x_n} = 0 \\ \lambda_1 \frac{\partial L}{\partial \lambda_1} &= 0, \dots, \lambda_m \frac{\partial L}{\partial \lambda_m} = 0\end{aligned}$$

Non-negativity Conditions

$$\begin{aligned}x_1 &\geq 0 \quad \dots \quad x_n \geq 0 \\ \lambda_1 &\geq 0 \quad \dots \quad \lambda_m \geq 0\end{aligned}$$

Note that some of these conditions are set equal to 0, while others are set as inequalities!

Note also that to minimize the function $f(x_1, \dots, x_n)$, the simplest thing to do is maximize the function $-f(x_1, \dots, x_n)$; all of the conditions remain the same after reformulating as a maximization problem.

There are additional assumptions (notably, $f(x)$ is quasi-concave and the constraints are convex) that are sufficient to ensure that a point satisfying the Kuhn-Tucker conditions is a global max; if these assumptions do not hold, you may have to check more than one point.

Finding the Critical Points with Kuhn-Tucker Conditions: Given the above conditions, to find the critical points we solve the above system of equations. To do so, we must check *all* border and interior solutions to see if they satisfy the above conditions.

In a two-dimensional set-up, this means we must check the following cases:

1. $x_1 = 0, x_2 = 0$ Border Solution
2. $x_1 = 0, x_2 \neq 0$ Border Solution
3. $x_1 \neq 0, x_2 = 0$ Border Solution

4. $x_1 \neq 0, x_2 \neq 0$ Interior Solution

Example 6.8 (Kuhn-Tucker with two variables). Solve the following optimization problem with inequality constraints

$$\max_{x_1, x_2} f(x) = -(x_1^2 + 2x_2^2)$$

$$\text{s.t.} \quad \begin{cases} x_1 + x_2 \leq 4 \\ x_1 \geq 0 \\ x_2 \geq 0 \end{cases}$$

Solution. Solve using the following steps:

1. Write the Lagrangian:

$$L(x_1, x_2, \lambda) = -(x_1^2 + 2x_2^2) - \lambda(x_1 + x_2 - 4)$$

2. Find the First Order Conditions: Kuhn-Tucker Conditions

$$\frac{\partial L}{\partial x_1} = -2x_1 - \lambda \leq 0$$

$$\frac{\partial L}{\partial x_2} = -4x_2 - \lambda \leq 0$$

$$\frac{\partial L}{\partial \lambda} = -(x_1 + x_2 - 4) \geq 0$$

Complementary Slackness Conditions

$$x_1 \frac{\partial L}{\partial x_2} = x_1(-2x_1 - \lambda) = 0$$

$$x_2 \frac{\partial L}{\partial x_2} = x_2(-4x_2 - \lambda) = 0$$

$$\lambda \frac{\partial L}{\partial \lambda} = -\lambda(x_1 + x_2 - 4) = 0$$

Non-negativity Conditions

$$x_1 \geq 0$$

$$x_2 \geq 0$$

$$\lambda \geq 0$$

3. Consider all border and interior cases:

Hypothesis	λ	x_1	x_2	$f(x_1, x_2)$
$x_1 = 0, x_2 = 0$	0	0	0	0
$x_1 = 0, x_2 \neq 0$	-16	0	4	-32
$x_1 \neq 0, x_2 = 0$	-8	4	0	-16
$x_1 \neq 0, x_2 \neq 0$	$-\frac{16}{3}$	$\frac{8}{3}$	$\frac{4}{3}$	$-\frac{32}{3}$

4. Find Maximum: Three of the critical points violate the requirement that $\lambda \geq 0$, so the point $(0, 0, 0)$ is the maximum.

◇

Exercise 6.2 (Kuhn-Tucker with logs). Solve the constrained optimization problem,

$$\max_{x_1, x_2} f(x) = \frac{1}{3} \log(x_1 + 1) + \frac{2}{3} \log(x_2 + 1) \text{ s.t. } \begin{array}{l} x_1 + 2x_2 \leq 4 \\ x_1 \geq 0 \\ x_2 \geq 0 \end{array}$$

6.8 Applications of Quadratic Forms

Curvature and The Taylor Polynomial as a Quadratic Form:

The Hessian is used in a Taylor polynomial approximation to $f(\mathbf{x})$ and provides information about the curvature of $f(\mathbf{x})$ at \mathbf{x} — e.g., which tells us whether a critical point \mathbf{x}^* is a min, max, or saddle point.

1. The second order Taylor polynomial about the critical point \mathbf{x}^* is

$$f(\mathbf{x}^* + \mathbf{h}) \approx \mathbf{f}(\mathbf{x}^*) + \nabla \mathbf{f}(\mathbf{x}^*) \mathbf{h} + \frac{1}{2} \mathbf{h}^\top \mathbf{H}(\mathbf{x}^*) \mathbf{h} + \mathbf{R}(\mathbf{h})$$

2. Since we're looking at a critical point, $\nabla f(\mathbf{x}^*) = 0$; and for small \mathbf{h} , $\mathbf{R}(\mathbf{h})$ is negligible. Rearranging, we get

$$f(\mathbf{x}^* + \mathbf{h}) - \mathbf{f}(\mathbf{x}^*) \approx \frac{1}{2} \mathbf{h}^\top \mathbf{H}(\mathbf{x}^*) \mathbf{h}$$

3. The RHS here is a quadratic form and we can determine the definiteness of $\mathbf{H}(\mathbf{x}^*)$.

Chapter 7

Probability Theory

Probability and Inferences are mirror images of each other, and both are integral to data science. Probability quantifies uncertainty, which is important because many things in the social world are at first uncertain. Inference is then the study of how to learn about facts you don't observe from facts you do observe.

7.1 Counting rules

7.1.1 Fundamental Theorem of Counting

If an object has j different characteristics that are independent of each other, and each characteristic i has n_i ways of being expressed, then there are $\prod_{i=1}^j n_i$ possible unique objects.

Example 7.1. Cards can be either red or black and can take on any of 13 values.

$$\begin{aligned} j &= \\ n_{\text{color}} &= \\ n_{\text{number}} &= \\ \text{Number of Outcomes} &= \end{aligned}$$

◇

We often need to count the number of ways to choose a subset from some set of possibilities. The number of outcomes depends on two characteristics of the process: does the order matter and is replacement allowed?

7.1.2 Sampling Table

If there are n objects which are numbered 1 to n and we select $k < n$ of them, how many different outcomes are possible?

	Order Matters	Order Doesn't Matter
With Replacement	n^k	$\binom{n+k-1}{k}$
Without Replacement	$n(n-1)(n-2)\dots(n-k+1) = \frac{n!}{(n-k)!}$	$\binom{n}{k} = \frac{n!}{(n-k)!k!}$

If the order in which a given object is selected matters, selecting 4 numbered objects in the following order (1, 3, 7, 2) and selecting the same four objects but in a different order such as (7, 2, 1, 3) will be counted as different outcomes.

If replacement is allowed, there are always the same n objects to select from. However, if replacement is not allowed, there is always one less option than the previous round when making a selection. For example, if replacement is not allowed and I am selecting 3 elements from the following set $\{1, 2, 3, 4, 5, 6\}$, I will have 6 options at first, 5 options as I make my second selection, and 4 options as I make my third.

Expression $\binom{n}{k}$ is read as “n choose k” and denotes $\frac{n!}{(n-k)!k!}$. Also, note that $0! = 1$.

Example 7.2 (Counting). There are five balls numbered from 1 through 5 in a jar. Three balls are chosen. How many possible choices are there?

1. Ordered, with replacement =
2. Ordered, without replacement =
3. Unordered, without replacement =

◇

Exercise 7.1 (Counting). Four cards are selected from a deck of 52 cards. Once a card has been drawn, it is not reshuffled back into the deck. Moreover, we care only about the complete hand that we get (i.e. we care about the set of selected cards, not the sequence in which it was drawn). How many possible outcomes are there?

◇

7.2 Sets

7.2.1 Set

A set is any well defined collection of elements. If x is an element of S , $x \in S$.

7.2.2 Sample Space (S)

A set or collection of all possible outcomes from some process. Outcomes in the set can be discrete elements (countable) or points along a continuous interval (uncountable).

Examples:

1. Discrete: the numbers on a die, number of days in a year.
2. Continuous: temperature, age.

7.2.3 Event

Any collection of possible outcomes of an experiment. Any subset of the full set of possibilities, including the full set itself. Event $A \subset S$.

7.2.4 Empty Set

a set with no elements. $S = \{\}$. It is denoted by the symbol \emptyset .

7.2.5 Set operations

1. **Union:** The union of two sets A and B , $A \cup B$, is the set containing all of the elements in A or B .

$$A_1 \cup A_2 \cup \cdots \cup A_n = \bigcup_{i=1}^n A_i$$

2. **Intersection:** The intersection of sets A and B , $A \cap B$, is the set containing all of the elements in both A and B .

$$A_1 \cap A_2 \cap \cdots \cap A_n = \bigcap_{i=1}^n A_i$$

3. **Complement:** If set A is a subset of S , then the complement of A , denoted A^C , is the set containing all of the elements in S that are not in A .

7.2.6 Properties of set operations

- **Commutative:** $A \cup B = B \cup A$; $A \cap B = B \cap A$
- **Associative:** $A \cup (B \cup C) = (A \cup B) \cup C$; $A \cap (B \cap C) = (A \cap B) \cap C$
- **Distributive:** $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$; $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- **de Morgan's laws:** $(A \cup B)^C = A^C \cap B^C$; $(A \cap B)^C = A^C \cup B^C$

- **Disjointness:** Sets are disjoint when they do not intersect, such that $A \cap B = \emptyset$. A collection of sets is pairwise disjoint (**mutually exclusive**) if, for all $i \neq j$, $A_i \cap A_j = \emptyset$. A collection of sets form a partition of set S if they are pairwise disjoint and they cover set S , such that $\bigcup_{i=1}^k A_i = S$.

Example 7.3 (Sets). Let set A be $\{1, 2, 3, 4\}$, B be $\{3, 4, 5, 6\}$, and C be $\{5, 6, 7, 8\}$. Sets A , B , and C are all subsets of the sample space S which is $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Write out the following sets:

1. $A \cup B$
2. $C \cap B$
3. B^c
4. $A \cap (B \cup C)$

◇

Exercise 7.2 (Sets). Suppose you had a pair of four-sided dice. You sum the results from a single toss. What is the set of possible outcomes (i.e. the sample space)? Consider subsets $A = \{2, 8\}$ and $B = \{2, 3, 7\}$ of the sample space you found. What is

1. A^c
2. $(A \cup B)^c$

◇

7.3 Probability

7.3.1 Probability Definitions: Formal and Informal

Many events or outcomes are random. In everyday speech, we say that we are *uncertain* about the outcome of random events. Probability is a formal model of uncertainty which provides a measure of uncertainty governed by a particular set of rules. A different model of uncertainty would, of course, have a different set of rules and measures. Our focus on probability is justified because it has proven to be a particularly useful model of uncertainty.

7.3.2 Probability Distribution Function

The Probability Distribution Function (PDF) is a mapping of each event in the sample space S to the real numbers that satisfy the following three axioms (also called Kolmogorov's Axioms).

Definition 7.1 (Probability). *Probability is a function that maps events to a real number, obeying the axioms of probability.* ♠

The axioms of probability make sure that the separate events add up in terms of probability, and – for standardization or *normalization* purposes – that sum of the probability events add up to 1.

Definition 7.2 (Axioms of Probability).

1. For any event A , $P(A) \geq 0$.
2. $P(S) = 1$
3. *The Countable Additivity Axiom: For any sequence of disjoint (mutually exclusive) events A_1, A_2, \dots (of which there may be infinitely many),*

$$P\left(\bigcup_{i=1}^k A_i\right) = \sum_{i=1}^k P(A_i)$$

The last axiom is an extension of a union to infinite sets. When there are only two events in the space, it boils down to:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) \quad \text{for disjoint } A_1, A_2$$

♠

7.3.3 Probability Operations

Using these three axioms, we can define all of the common rules of probability.

1. $P(\emptyset) = 0$
2. For any event A , $0 \leq P(A) \leq 1$.
3. $P(A^C) = 1 - P(A)$
4. If $A \subset B$ (A is a subset of B), then $P(A) \leq P(B)$.
5. For *any* two events A and B , $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
6. Boole's Inequality: For any sequence of n events (which need not be disjoint) A_1, A_2, \dots, A_n , then $P\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n P(A_i)$.

Example 7.4 (Probability). Let's assume we have an evenly-balanced, six-sided die. Then,

1. Sample space $S =$
2. $P(1) = \dots = P(6) =$
3. $P(\emptyset) = P(7) =$
4. $P(\{1, 3, 5\}) =$
5. $P(\{1, 2\}^C) = P(\{3, 4, 5, 6\}) =$
6. Let $A = \{1, 2, 3, 4, 5\} \subset S$. Then $P(A) = 5/6 < P(S) =$
7. Let $A = \{1, 2, 3\}$ and $B = \{2, 4, 6\}$. Then $A \cup B$? $A \cap B$? $P(A \cup B)$?

◇

Exercise 7.3 (Probability). Suppose you had a pair of four-sided dice. You sum the results from a single toss. Let us call this sum, or the outcome, X .

1. What is $P(X = 5)$, $P(X = 3)$, $P(X = 6)$?
2. What is $P(X = 5 \cup X = 3)^C$?

◇

7.4 Conditional Probability and Bayes Law

7.4.1 Conditional Probability

The conditional probability $P(A|B)$ of an event A is the probability of A , given that another event B has occurred (read: “probability of A given B ”). Conditional probability allows for the inclusion of other information into the calculation of the probability of an event. It is calculated as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Note that conditional probabilities, are probabilities that also follow the Kolmagorov axioms of probability.

Example 7.5 (Conditional Probability 1). Assume A and B occur with the following frequencies:

	A	A^c
B	n_{ab}	$n_{a^c b}$
B^c	n_{ab^c}	$n_{(ab)^c}$

and let $n_{ab} + n_{a^c b} + n_{ab^c} + n_{(ab)^c} = N$. Then

1. $P(A) =$
2. $P(B) =$
3. $P(A \cap B) =$
4. $P(A|B) = \frac{P(A \cap B)}{P(B)} =$
5. $P(B|A) = \frac{P(A \cap B)}{P(A)} =$

◇

Example 7.6 (Conditional Probability 2). A six-sided die is rolled. What is the probability of a 1, given the outcome is an odd number?

◇

You could rearrange the fraction to highlight how a joint probability could be expressed as the product of a conditional probability.

Definition 7.3 (Multiplicative Law of Probability). *The probability of the intersection of two events A and B is $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$ which follows directly from the definition of conditional probability. More generally,*

$$\begin{aligned} P(A_1 \cap \cdots \cap A_k) \\ &= P(A_k | A_{k-1} \cap \cdots \cap A_1) \times P(A_{k-1} | A_{k-2} \cap \cdots \cap A_1) \times \cdots \\ &\quad \times P(A_2 | A_1) \times P(A_1) \end{aligned}$$

Sometimes it is easier to calculate these conditional probabilities and sum them than it is to calculate $P(A)$ directly. ♠

Definition 7.4 (Law of Total Probability). *Let S be the sample space of some experiment and let the disjoint k events B_1, \dots, B_k partition S , such that $P(B_1 \cup \dots \cup B_k) = P(S) = 1$. If A is some other event in S , then the events $A \cap B_1, A \cap B_2, \dots, A \cap B_k$ will form a partition of A and we can write A as*

$$A = (A \cap B_1) \cup \cdots \cup (A \cap B_k).$$

Since the k events are disjoint,

$$\begin{aligned} P(A) &= \sum_{i=1}^k P(A \cap B_i) \\ &= \sum_{i=1}^k P(B_i)P(A|B_i). \end{aligned}$$

♠

7.4.2 Bayes Rule

Assume that events B_1, \dots, B_k form a partition of the space S . Then by the Law of Total Probability

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(B_j)P(A|B_j)}{\sum_{i=1}^k P(B_i)P(A|B_i)}$$

If there are only two states of B , then this is just

$$P(B_1|A) = \frac{P(B_1)P(A|B_1)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2)}$$

Bayes' rule determines the posterior probability of a state $P(B_j|A)$ by calculating the probability $P(A \cap B_j)$ that both the event A and the state B_j will occur and dividing it by the probability that the event will occur regardless of the state (by summing across all B_i). The states could be something like Normal/Defective, Healthy/Diseased, etc.. The event on which one conditions could be something like a sampling from a batch of components, or a test for a disease.

7.4.3 Prior and Posterior Probabilities

Above, $P(B_1)$ is often called the prior probability, since it's the probability of B_1 before anything else is known. $P(B_1|A)$ is called the posterior probability, since it's the probability after other information is taken into account.

Example 7.7 (Bayes' Rule). A test for cancer correctly detects cancer 90% of the time, and incorrectly 10% of the time. If 10% of the population have cancer at any given time, what is the probability that a person who tests positive actually has cancer?

◇

Exercise 7.4 (Conditional Probability). Assume that 2% of the population like the color green. We develop a survey that positively classifies someone as liking green given that they wear green 95% of the time and negatively classifies someone as not liking green given that they wear another color clothing 97% of the time. What is the probability that someone positively classified as green, actually likes green?

◇

7.5 Independence

Definition 7.5 (Independence). *If the occurrence or non-occurrence of either events A and B have no effect on the occurrence or non-occurrence of the other, then A and B are independent.* ♠

If A and B are independent, then

1. $P(A|B) = P(A)$
2. $P(B|A) = P(B)$
3. $P(A \cap B) = P(A)P(B)$
4. More generally than the above, $P(\bigcap_{i=1}^k A_i) = \prod_{i=1}^k P(A_i)$

Are mutually exclusive events independent of each other? No. If A and B are mutually exclusive, then they cannot happen simultaneously. If we know that A occurred, then we know that B couldn't have occurred. Because of this, A and B aren't independent.

7.5.1 Pairwise Independence

A set of more than two events A_1, A_2, \dots, A_k is pairwise independent if $P(A_i \cap A_j) = P(A_i)P(A_j)$, $\forall i \neq j$. Note that this does **not** necessarily imply joint independence.

7.5.2 Conditional Independence

If A and B are independent once you know the occurrence of a third event C , then A and B are conditionally independent (conditional on C):

1. $P(A|B \cap C) = P(A|C)$
2. $P(B|A \cap C) = P(B|C)$
3. $P(A \cap B|C) = P(A|C)P(B|C)$

Just because two events are conditionally independent does not mean that they are independent. Actually it is hard to think of real-world things that are “unconditionally” independent. That’s why it’s always important to ask about a finding: What was it conditioned on? For example, suppose that a company hiring decisions are done by only one manager, who picks a group of 50 bright candidates, and flips a coin for each candidate to generate a team of about 25 employees. Then the the probability that two employees get accepted are conditionally independent, because they are determined by two separate coin tosses. However, this does not mean that their admittance is not completely independent. Knowing that employee A is accepted gives us information about whether employee B is accepted, if we think that the manager originally picked her pool of 50 candidates by merit.

Perhaps more counter-intuitively: If two events are already independent, then it might seem that no amount of “conditioning” will make

them dependent. But this is not always so. For example, suppose I only get a call from two people, Alice and Bob. Let A be the event that Alice calls, and B be the event that Bob calls. Alice and Bob do not communicate, so $P(A | B) = P(A)$. But now let C be the event that your phone rings. For conditional independence to hold here, then $P(A | C)$ must be equal to $P(A | B \cap C)$. But this is not true – $A | C$ may or may not be true, but $P(A | B \cap C)$ certainly is true.

7.6 Random Variables

Most questions in data science involve events, rather than numbers per se. To analyze and reason about events quantitatively, we need a way of mapping events to numbers. A random variable does exactly that.

Definition 7.6 (Random Variable). *A random variable is a measurable function X that maps from the sample space S to the set of real numbers R . It assigns a real number to every outcome $s \in S$.* ♠

It might seem strange to define a random variable as a function – which is neither random nor variable. The randomness comes from the realization of an event from the sample space s .

7.6.1 Randomness

means that the outcome of some experiment is not deterministic, i.e. there is some probability ($0 < P(A) < 1$) that the event will occur.

The support of a random variable is all values for which there is a positive probability of occurrence.

Example: Flip a fair coin two times. What is the sample space?

A random variable must map events to the real line. For example, let a random variable X be the number of heads. The event (H, H) gets mapped to 2 ($X(s) = 2$), and the events $\{(H, T), (T, H)\}$ gets mapped to 1 ($X(s) = 1$), the event (T, T) gets mapped to 0 ($X(s) = 0$).

What are other possible random variables?

7.7 Distributions

We now have two main concepts in this section – probability and random variables. Given a sample space S and the same experiment, both probability and random variables take events as their inputs. But they output different things (probabilities measure the “size” of events, random variables give a number in a way that the analyst chose to define the random variable). How do the two concepts relate?

The concept of distributions is the natural bridge between these two concepts.

Definition 7.7 (Distribution of a random variable). *A distribution of a random variable is a function that specifies the probabilities of all events associated with that random variable. There are several types of distributions: A probability mass function for a discrete random variable and probability density function for a continuous random variable. ♠*

Notice how the definition of distributions combines two ideas of random variables and probabilities of events. First, the distribution considers a random variable, call it X . X can take a number of possible numeric values.

Example 7.8 (Total Number of Occurrences). Consider three binary outcomes, one for each patient recovering from a disease: R_i denotes the event in which patient i ($i = 1, 2, 3$) recovers from a disease. R_1 , R_2 , and R_3 . How would we represent the total number of people who end up recovering from the disease?

Solution. Define the random variable X be the total number of people (out of three) who recover from the disease. Random variables are functions, that take as an input a set of events (in the sample space S) and deterministically assigns them to a number of the analyst's choice.

◇

Recall that with each of these numerical values there is a class of *events*. In the previous example, for $X = 3$ there is one outcome (R_1, R_2, R_3) and for $X = 1$ there are multiple $(\{(R_1, R_2^c, R_3^c), (R_1^c, R_2, R_3^c), (R_1^c, R_2^c, R_3), \})$. Now, the thing to notice here is that each of these events naturally come with a probability associated with them. That is, $P(R_1, R_2, R_3)$ is a number from 0 to 1, as is $P(R_1, R_2^c, R_3^c)$. These all have probabilities because they are in the sample space S . The function that tells us these probabilities that are associated with a numerical value of a random variable is called a distribution.

In other words, a random variable X *induces a probability distribution* P (sometimes written P_X to emphasize that the probability density is about the r.v. X)

7.7.1 Discrete Random Variables

The formal definition of a random variable is easier to given by separating out two cases: discrete random variables when the numeric summaries of the events are discrete, and continuous random variables when they are continuous.

Definition 7.8 (Discrete Random Variable). X is a discrete random variable if it can assume only a finite or countably infinite number of distinct values. Examples: number of wars per year, heads or tails.



The distribution of a discrete r.v. is a PMF:

Definition 7.9 (Probability Mass Function). For a discrete random variable X , the probability mass function (PMF) – also referred to simply as the “probability distribution” – $p(x) = P(X = x)$, assigns probabilities to a countable number of distinct x values such that

1. $0 \leq p(x) \leq 1$
2. $\sum_y p(x) = 1$



Example: For a fair six-sided die, there is an equal probability of rolling any number. Since there are six sides, the probability mass function is then $p(y) = 1/6$ for $y = 1, \dots, 6$, 0 otherwise.}

In a discrete random variable, **cumulative density function** (CDF) – also referred to simply as the “cumulative distribution” or previously as the “density function” – $F(x)$ or $P(X \leq x)$, is the probability that X is less than or equal to some value x , or

$$P(X \leq x) = \sum_{i \leq x} p(i)$$

Properties a CDF must satisfy:

1. $F(x)$ is non-decreasing in x .
2. $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$
3. $F(x)$ is right-continuous.

Note that $P(X > x) = 1 - P(X \leq x)$.

Example 7.9. For a fair die with its value as Y , what are the following?

1. $P(Y \leq 1) \implies P(Y \leq 1) = \frac{1}{6}$
2. $P(Y \leq 3) \implies P(Y \leq 3) = \frac{1}{2}$
3. $P(Y \leq 6) \implies P(Y \leq 6) = 1$



7.7.2 Continuous Random Variables

We also have a similar definition for *continuous* random variables.

Definition 7.10 (Continuous Random Variable). *X is a continuous random variable if there exists a nonnegative function $f(x)$ defined for all real $x \in (-\infty, \infty)$, such that for any interval A , $P(X \in A) = \int_A f(x)dx$. Examples: age, income, GNP, temperature.* ♠

Definition 7.11 (Probability Density Function). *The function f above is called the probability density function (PDF) of X and must satisfy*

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Note also that $P(X = x) = 0$ — i.e., the probability of any point y is zero. ♠

Example: $f(y) = 1, \quad 0 \leq y \leq 1$

For both discrete and continuous random variables, we have a unifying concept of another measure: the cumulative distribution:

Definition 7.12 (Cumulative Density Function). *Because the probability that a continuous random variable will assume any particular value is zero, we can only make statements about the probability of a continuous random variable being within an interval. The cumulative distribution gives the probability that Y lies on the interval $(-\infty, y)$ and is defined as*

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(s)ds.$$

Note that $F(x)$ has similar properties with continuous distributions as it does with discrete - non-decreasing, continuous (not just right-continuous), and $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$. ♠

We can also make statements about the probability of Y falling in an interval $a \leq y \leq b$.

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$

The PDF and CDF are linked by the integral. The CDF of the integral of the PDF:

$$f(x) = F'(x) = \frac{dF(x)}{dx}$$

Example 7.10. For $f(y) = 1$, $0 < y < 1$, find:

1. The CDF $F(y) \implies F(y) = \int_0^y f(s)ds = \int_0^y 1ds = s \Big|_0^y = y$.
2. The probability $P(0.5 < y < 0.75) \implies \int_{0.5}^{0.75} 1ds = s \Big|_{0.5}^{0.75} = 0.25$.

◇

7.8 Joint Distributions

Often, we are interested in two or more random variables defined on the same sample space. The distribution of these variables is called a joint distribution. Joint distributions can be made up of any combination of discrete and continuous random variables.

Joint Probability Distribution: If both X and Y are random variable, their joint probability mass/density function assigns probabilities to each pair of outcomes

Discrete:

$$p(x, y) = P(X = x, Y = y)$$

such that $p(x, y) \in [0, 1]$ and

$$\sum \sum p(x, y) = 1$$

Continuous:

$$f(x, y); P((X, Y) \in A) = \iint_A f(x, y) dx dy$$

s.t. $f(x, y) \geq 0$ and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$$

If X and Y are independent, then $P(X = x, Y = y) = P(X = x)P(Y = y)$ and $f(x, y) = f(x)f(y)$

7.8.1 Marginal Probability Distribution

The Marginal Probability Distribution is a probability distribution of only one of the two variables (ignoring information about the other variable). We can obtain the marginal distribution by summing/integrating across the variable that we don't care about:

- Discrete: $p_X(x) = \sum_i p(x, y_i)$
- Continuous: $f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$

7.8.2 Conditional Probability Distribution

The Conditional Probability Distribution is a probability distribution for one variable, holding the other variable fixed. Recalling from the previous lecture that $P(A|B) = \frac{P(A \cap B)}{P(B)}$, we can write the conditional distribution as

- Discrete: $p_{Y|X}(y|x) = \frac{p(x,y)}{p_X(x)}$, $p_X(x) > 0$
- Continuous: $f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$, $f_X(x) > 0$

Exercise 7.5 (Discrete Outcomes). Suppose we are interested in the outcomes of flipping a coin and rolling a 6-sided die at the same time. The sample space for this process contains 12 elements:

$$\{(H, 1), (H, 2), (H, 3), (H, 4), (H, 5), (H, 6), \\ (T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6)\}$$

We can define two random variables X and Y such that $X = 1$ if heads and $X = 0$ if tails, while Y equals the number on the die.

We can then make statements about the joint distribution of X and Y . What are the following?

1. $P(X = x)$
2. $P(Y = y)$
3. $P(X = x, Y = y)$
4. $P(X = x|Y = y)$
5. Are X and Y independent?

◇

7.9 Expectation

We often want to summarize some characteristics of the distribution of a random variable. The most important summary is the expectation (or expected value, or mean), in which the possible values of a random variable are weighted by their probabilities.

7.9.1 Expected Value of a Discrete Random Variable

Definition 7.13 (Expected Value of a Discrete Random Variable). *The expected value of a discrete random variable Y is*

$$E(Y) = \sum_y yP(Y = y) = \sum_y yp(y)$$

In words, it is the weighted average of all possible values of Y , weighted by the probability that y occurs. It is not necessarily the number we would expect Y to take on, but the average value of Y after a large number of repetitions of an experiment. ♠

Example 7.11 (Expected Value of a Discrete Random Variable). What is the expectation of a fair, six-sided die?

$$\Rightarrow E(Y) = \sum_{y=1}^6 yp(y) = \frac{1}{6} \sum_{y=1}^6 y = \frac{1}{6} + 2\frac{1}{6} + \cdots + 6\frac{1}{6} = \frac{21}{6} = 3.5$$

◇

7.9.2 Expected Value of a Continuous Random Variable

The expected value of a continuous random variable is similar in concept to that of the discrete random variable, except that instead of summing using probabilities as weights, we integrate using the density to weight. Hence, the expected value of the continuous variable Y is defined by

$$E(Y) = \int_y yf(y)dy$$

Example 7.12 (Expected Value of a Continuous Random Variable). Find $E(Y)$ for $f(y) = \frac{1}{1.5}$, $0 < y < 1.5$.

$$\Rightarrow E(Y) = \int_0^{1.5} \frac{2}{3}y \, dy = \frac{1}{3}y^2 \Big|_0^{1.5} = 0.75$$

◇

7.9.3 Expected Value of a Function

Remember: An Expected Value is a type of weighted average. We can extend this to composite functions. For random variable Y , if Y is Discrete with PMF $p(y)$,

$$E[g(Y)] = \sum_y g(y)p(y)$$

if Y is Continuous with PDF $f(y)$,

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy.$$

7.9.4 Properties of Expected Values

Dealing with the Expected Value is easier when the “thing” inside is a sum. The intuition behind this that Expectation is an integral, which is a type of sum.

1. Expectation of a constant is a constant

$$E(c) = c.$$

2. Constants come out

$$E(cg(Y)) = cE(g(Y)).$$

3. Expectation is Linear

$$E(g(Y_1) + \cdots + g(Y_n)) = E(g(Y_1)) + \cdots + E(g(Y_n)),$$

regardless of independence.

4. Expected Value of Expected Values:

$$E(E(Y)) = E(Y),$$

(because the expected value of a random variable is a constant).

Finally, if X and Y are independent, even products are easy:

$$E(XY) = E(X)E(Y).$$

7.9.5 Conditional Expectation

With joint distributions, we are often interested in the expected value of a variable Y if we could hold the other variable X fixed. This is the conditional expectation of Y given $X = x$:

1. Y discrete: $E(Y|X = x) = \sum_y yp_{Y|X}(y|x)$
2. Y continuous: $E(Y|X = x) = \int_y yf_{Y|X}(y|x)dy$

The conditional expectation is often used for prediction when one knows the value of X but not Y

7.10 Variance and Covariance

We can also look at other summaries of the distribution, which build on the idea of taking expectations. Variance tells us about the “spread” of the distribution; it is the expected value of the squared deviations from the mean of the distribution. The standard deviation is simply the square root of the variance.

Definition 7.14 (Variance). *The Variance of a Random Variable Y is*

$$\text{Var}(Y) = E[(Y - E(Y))^2] = E(Y^2) - [E(Y)]^2$$

The Standard Deviation is the square root of the variance:

$$SD(Y) = \sigma_Y = \sqrt{\text{Var}(Y)}$$



Example 7.13 (Variance). Given the following PMF:

$$f(x) = \begin{cases} \frac{3!}{x!(3-x)!} \left(\frac{1}{2}\right)^3 & x = 0, 1, 2, 3 \\ 0 & \text{otherwise} \end{cases}$$

What is $\text{Var}(x)$? **Hint:** First calculate $E(X)$ and $E(X^2)$



Definition 7.15 (Covariance). *The covariance measures the degree to which two random variables vary together; if the covariance between X and Y is positive, X tends to be larger than its mean when Y is larger than its mean.*

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

We can also write this as

$$\begin{aligned} \text{Cov}(X, Y) &= E(XY - XE(Y) - E(X)Y + E(X)E(Y)) \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$



The covariance of a variable with itself is the variance of that variable. The covariance is unfortunately hard to interpret in magnitude. The correlation is a standardized version of the covariance, and always ranges from -1 to 1.

Definition 7.16 (Correlation). *The correlation coefficient is the covariance divided by the standard deviations of X and Y . It is a unitless measure and always takes on values in the interval $[-1, 1]$.*

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{SD(X)SD(Y)}$$



7.10.1 Properties of Variance and Covariance

1. $\text{Var}(c) = 0$
2. $\text{Var}(cY) = c^2 \text{Var}(Y)$
3. $\text{Cov}(Y, Y) = \text{Var}(Y)$
4. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
5. $\text{Cov}(aX, bY) = ab \text{Cov}(X, Y)$
6. $\text{Cov}(X + a, Y) = \text{Cov}(X, Y)$
7. $\text{Cov}(X + Z, Y + W) = \text{Cov}(X, Y) + \text{Cov}(X, W) + \text{Cov}(Z, Y) + \text{Cov}(Z, W)$
8. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$

Exercise 7.6 (Expectation and Variance). Suppose we have a PMF with the following characteristics:

$$P(X = -2) = \frac{1}{5}$$

$$P(X = -1) = \frac{1}{6}$$

$$P(X = 0) = \frac{1}{5}$$

$$P(X = 1) = \frac{1}{15}$$

$$P(X = 2) = \frac{11}{30}$$

1. Calculate the expected value of X

Define the random variable $Y = X^2$.

2. Calculate the expected value of Y . (Hint: It would help to derive the PMF of Y first in order to calculate the expected value of Y in a straightforward way)

3. Calculate the variance of X .

◇

Exercise 7.7 (Expectation and Variance 2). Find the expectation and variance, given the following PDF:

$$f(x) = \begin{cases} \frac{3}{10}(3x - x^2) & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

◇

Exercise 7.8 (Expectation and Variance 3). Consider the following:

1. Find the mean and standard deviation of random variable X . The PDF of this X is as follows:

$$f(x) = \begin{cases} \frac{1}{4}x & 0 \leq x \leq 2 \\ \frac{1}{4}(4 - x) & 2 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

2. Next, calculate $P(X < \mu - \sigma)$. Remember, μ is the mean and σ is the standard deviation

◇

7.11 Special Distributions

7.11.1 Common Discrete distributions

Definition 7.17 (Binomial Distribution). Y is distributed binomial if it represents the number of “successes” observed in n independent, identical “trials,” where the probability of success in any trial is p and the probability of failure is $q = 1 - p$. ♠

For any particular sequence of y successes and $n - y$ failures, the probability of obtaining that sequence is $p^y q^{n-y}$ (by the multiplicative law and independence). However, there are $\binom{n}{y} = \frac{n!}{(n-y)!y!}$ ways of obtaining a sequence with y successes and $n - y$ failures. So the binomial distribution is given by

$$p(y) = \binom{n}{y} p^y q^{n-y}, \quad y = 0, 1, 2, \dots, n$$

with mean $\mu = E(Y) = np$ and variance $\sigma^2 = \text{Var}(Y) = npq$.

Example 7.14. Cats vote for Dog-sponsored bills 2% of the time. What is the probability that out of 10 Cats questioned, half voted for a particular Dog-sponsored bill? What is the mean number of Cats voting for Dog-sponsored bills? The variance? 1. $P(Y = 5) = 1$. $E(Y) = 1$. $\text{Var}(Y) = 6$

Solution. Let $n = 10$, $y = 5$, and $p = 0.02$.

- $P(Y = 5) = \binom{10}{5}(0.02)^5(1 - 0.02)^{10-5} = \binom{10}{5}(0.02)^5(0.93)^5 = 0.073\%$
- $E(Y) = pn = 0.2$
- $V(Y) = npq = 10 \times 0.02 \times 0.98 = 0.196$

◇

Definition 7.18 (Poisson Distribution). A random variable Y has a Poisson distribution if

$$P(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}, \quad y = 0, 1, 2, \dots, \quad \lambda > 0$$

The Poisson has the unusual feature that its expectation equals its variance: $E(Y) = \text{Var}(Y) = \lambda$. The Poisson distribution is often used to model rare event counts: counts of the number of events that occur during some unit of time. λ is often called the “arrival rate.” ♠

Example 7.15. Phone calls at a call center occur through a Poisson Distribution, at a rate of 2 per hour. What is the probability of 0, 2, and less than 5 calls occurring in a hour?

Solution. Given the above conditions, the probabilities are

- $P(H = 0) = P(0) = \frac{2^0}{0!} \exp^{-2} = 0.13$
- $P(H = 2) = P(2) = \frac{2^2}{2!} \exp^{-2} = 0.27$
- $P(H < 5) = \sum_{y=0}^4 \frac{2^y}{y!} \exp^{-2} = 0.95$

◇

7.11.2 Common Continuous Distributions

Definition 7.19 (Uniform Distribution). A random variable Y has a continuous uniform distribution on the interval (α, β) if its density is given by

$$f(y) = \frac{1}{\beta - \alpha}, \quad \alpha \leq y \leq \beta$$

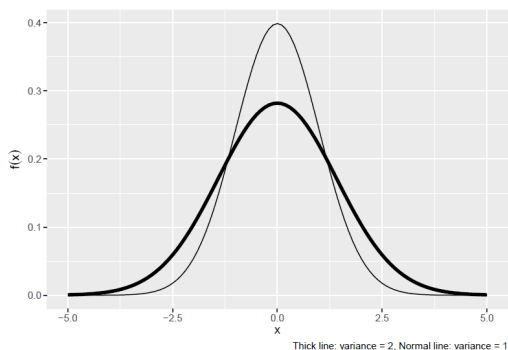


Figure 7.1: Normal Distribution Density

The mean and variance of Y are $E(Y) = \frac{\alpha+\beta}{2}$ and $\text{Var}(Y) = \frac{(\beta-\alpha)^2}{12}$. ♠

Example 7.16. For Y uniformly distributed over $(1, 3)$, what are the following probabilities?

1. $P(Y = 2) \implies P(Y = 2) = 0$
2. $f(2) \implies f(2) = \frac{1}{3-1} = \frac{1}{2}$
3. $P(Y \leq 2) \implies P(Y \leq 2) = F(2) = \int_1^2 \frac{1}{2} dy = \frac{1}{2}$
4. $P(Y > 2) \implies P(Y > 2) = 1 - P(Y \leq 2) = \frac{1}{2}$

◇

Definition 7.20 (Normal Distribution). A random variable Y is normally distributed with mean $E(Y) = \mu$ and variance $\text{Var}(Y) = \sigma^2$ if its density is

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

♠

See Figure 7.1 are various Normal Distributions with the same $\mu = 1$ and two versions of the variance.

7.12 Summarizing Observed Events (Data)

So far, we’ve talked about distributions in a theoretical sense, looking at different properties of random variables. We don’t observe random variables; we observe realizations of the random variable. These realizations of events are roughly equivalent to what we mean by “data”.

7.12.1 Sample mean

This is the most common measure of central tendency, calculated by summing across the observations and dividing by the number of observations.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

The sample mean is an *estimate* of the expected value of a distribution.

Example 7.17. Use the table below to calculate:

X	6	3	7	5	5	5	6	4	7	2
Y	1	2	1	2	2	1	2	0	2	0

1. $\bar{x} = \frac{50}{10} = 5$ $\bar{y} = \frac{13}{10} = 1.3$
2. $\text{median}(x) = 5$ $\text{median}(y) = 1.5$
3. $m_x = 5$ $m_y = 2$

◇

7.12.2 Dispersion

We also typically want to know how spread out the data are relative to the center of the observed distribution. There are several ways to measure dispersion.

7.12.3 Sample variance

The sample variance is the sum of the squared deviations from the sample mean, divided by the number of observations minus 1.

$$\hat{\text{Var}}(X) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Again, this is an *estimate* of the variance of a random variable; we divide by $n-1$ instead of n in order to get an unbiased estimate.

7.12.4 Standard deviation

The sample standard deviation is the square root of the sample variance.

$$\hat{SD}(X) = \sqrt{\hat{\text{Var}}(X)} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Example 7.18. Using table above, calculate:

1. $\text{Var}(X) = 2.67$ $\text{Var}(Y) = 0.68$
2. $\text{SD}(X) = 1.63$ $\text{SD}(Y) = 0.82$

◇

7.12.5 Covariance and Correlation

Both of these quantities measure the degree to which two variables vary together, and are estimates of the covariance and correlation of two random variables as defined above.

1. **Sample covariance:** $\hat{\text{Cov}}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
2. **Sample correlation:** $\hat{\text{Corr}} = \frac{\hat{\text{Cov}}(X, Y)}{\sqrt{\hat{\text{Var}}(X)\hat{\text{Var}}(Y)}}$

Example 7.19. Using the above table, calculate the sample versions of:

1. $\text{Cov}(X, Y) = 0.56$
2. $\text{Corr}(X, Y) = 0.41$

◇

7.13 Asymptotic Theory

In theoretical and applied research, asymptotic arguments are often made. In this section we briefly introduce some of this material.

What are asymptotics? In probability theory, asymptotic analysis is the study of limiting behavior. By limiting behavior, we mean the behavior of some random process as the number of observations gets larger and larger.

Why is this important? We rarely know the true process governing the events we see in the natural world. It is helpful to understand how such unknown processes theoretically must behave and asymptotic theory helps us do this.

7.13.1 CLT and LLN

We are now finally ready to revisit, with a bit more precise terms, the two pillars of statistical theory we motivated Section 4.

Theorem 7.9 (Central Limit Theorem (i.i.d. case)). *Let $\{X_n\} = \{X_1, X_2, \dots\}$ be a sequence of i.i.d. random variables with finite mean (μ) and variance (σ^2) . Then, the sample mean $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$ increasingly converges into a Normal distribution.*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \text{Normal}(0, 1),$$

Another way to write this as a probability statement is that for all real numbers a ,

$$P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq a\right) \rightarrow \Phi(a)$$

as $n \rightarrow \infty$, where

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

is the CDF of a Normal distribution with mean 0 and variance 1.

This result means that, as n grows, the distribution of the sample mean $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$ is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, i.e.,

$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

The standard deviation of \bar{X}_n (which is roughly a measure of the precision of \bar{X}_n as an estimator of μ) decreases at the rate $1/\sqrt{n}$, so, for example, to increase its precision by 10 (i.e., to get one more digit right), one needs to collect $10^2 = 100$ times more units of data.

Intuitively, this result also justifies that whenever a lot of small, independent processes somehow combine together to form the realized observations, practitioners often feel comfortable assuming Normality. ♠

Theorem 7.10 (Law of Large Numbers (LLN)). *For any draw of independent random variables with the same mean μ , the sample average after n draws, $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$, converges in probability to the expected value of X , μ as $n \rightarrow \infty$:*

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$$

A shorthand of which is $\bar{X}_n \xrightarrow{p} \mu$, where the arrow is read as “converges in probability to” as $n \rightarrow \infty$. ♠

In other words, $P(\lim_{n \rightarrow \infty} \bar{X}_n = \mu) = 1$. This is an important motivation for the widespread use of the sample mean, as well as the intuition link between averages and expected values.

More precisely this version of the LLN is called the *weak* law of large numbers because it leaves open the possibility that $|\bar{X}_n - \mu| > \varepsilon$ occurs many times. The *strong* law of large numbers states that, under a few more conditions, the probability that the limit of the sample average is the true mean is 1 (and other possibilities occur with probability 0), but the difference is rarely consequential in practice.

The Strong Law of Large Numbers holds so long as the expected value exists; no other assumptions are needed. However, the rate of convergence will differ greatly depending on the distribution underlying the observed data. When extreme observations occur often (i.e. kurtosis is large), the rate of convergence is much slower. Cf. The distribution of financial returns.

7.13.2 Big O-Notation

Some of you may encounter “big-O”-notation. If f, g are two functions, we say that $f = \mathcal{O}(g)$ if there exists some constant, c , such that $f(n) \leq c \times g(n)$ for large enough n . This notation is useful for simplifying complex problems in game theory, computer science, and statistics.

Example 7.20. What is $\mathcal{O}(5 \exp(0.5n) + n^2 + n/2)$?

Solution. $\exp(n)$. Why? Because, for large n ,

$$\frac{5 \exp(0.5n) + n^2 + n/2}{\exp(n)} \leq \frac{c \exp(n)}{\exp(n)} = c.$$

whenever $n > 4$ and where $c = 1$.

◇

Appendix A

Conventions & Symbols

A.1 Notation

Tables A.1-A.5 summarize the notational conventions that are used throughout this text. A non-bold face symbol denotes a scalar quantity. A bold face symbol denotes either a vector (typically lower case) or a matrix (typically upper case). It is important to make the distinction between a *true* value, a *calculated*, *estimated*, or a *measured* value. As shown in Table A.1, the true value has no additional mark; the calculated value has a “hat” on it; the measured value has a “tilde” above it. The error is defined as the true value minus the estimated value. The error quantity is indicated with a δ , for example $\delta\mathbf{x} = \mathbf{x} - \hat{\mathbf{x}}$.

A.2 Acronyms

The acronyms used in the text should be defined at their first usage which should be listed in the index.

A.3 Greek letters

The Greek letters used in the text are defined in Table A.6, with their proper pronunciation.

Table A.1: Notational conventions.

x	non-bold face variables denote <i>scalars</i>
\mathbf{x}	boldface lower-case denotes <i>vector</i> quantities
\mathbf{X}	boldface upper-case denotes <i>matrix</i> quantities
$x_{i,j}$	row i and column j entry of matrix \mathbf{X}
\mathbf{x}	true value of \mathbf{x}
$\hat{\mathbf{x}}$	calculated value of \mathbf{x}
$\tilde{\mathbf{x}}$	measured value of \mathbf{x}
$\delta\mathbf{x}$	error $\mathbf{x} - \hat{\mathbf{x}}$
\mathbf{R}_a^b	transformation matrix from reference frames a to b
\mathbf{x}^a	vector \mathbf{x} represented with respect to frame a
$\mathbb{R}, \mathbb{R}^+, \mathbb{R}^n$	real numbers, reals greater than 0, n -tuples of reals
\mathbb{N}	natural numbers $\{0, 1, 2, \dots\}$
\mathbb{C}	complex numbers
\mathbb{Z}	integer numbers
$\mathbf{0}_{n \times m}$ or $\mathbf{0}$	zero matrix
$\mathbf{I}_{n \times n}$ or \mathbf{I}	identity matrix
$ \mathbf{X} $	determinant of matrix \mathbf{X}
R, N	range space, null space
R_∞, N_∞	generalized range space and null space
\mathcal{N}	Normal or Gaussian random variable
\mathcal{L}	Laplace random variable
■	end of proof, "I have proved"
◇	end of example or exercise
♠	end of theorem or definition

Table A.2: Equivalence symbols.

$=$	equal to
\neq	not equal to
$>$	greater than
$<$	less than
\geq	greater than or equal to
\leq	less than or equal to
\propto	proportional to
\approx	approximately equal to
\sim	distributed as (or indifference)
\equiv	equivalent to
\triangleq	computed as
\succ	preferred to

Table A.3: Set notation symbols.

$(a .. b), [a .. b]$	open interval, closed interval
$\langle \dots \rangle$	sequence (a list in which order matters)
$\{ \dots \}$	set (a list in which order does not matter)
\in	is an element of
\emptyset	empty set
\cup	union
\cap	intersection
\subset	subset

Table A.4: Logical symbols.

\therefore	therefore
\forall	for all
\exists	there exists
\implies	logical "then" statement
\iff	if and only if

Table A.5: Abbreviations.

<i>iff</i>	if an only if
s.t.	such that
LHS	left hand side
RHS	right hand side
QED	end of proof, "I have proved"
w.r.t.	with respect to

Table A.6: Greek letters with pronunciation.

α	alpha <i>AL-fuh</i>
β	beta <i>BAY-tuh</i>
γ, Γ	gamma <i>GAM-muh</i>
δ, Δ	delta <i>DEL-tuh</i>
ϵ	epsilon <i>EP-suh-lon</i>
ζ	zeta <i>ZAY-tuh</i>
η	eta <i>AY-tuh</i>
θ, Θ	theta <i>THAY-tuh</i>
ι	iota <i>eye-OH-tuh</i>
κ	kappa <i>KAP-uh</i>
λ, Λ	lambda <i>LAM-duh</i>
μ	mu <i>MEW</i>
ν	nu <i>NEW</i>
ξ, Ξ	xi <i>KSIGH</i>
\omicron	omicron <i>OM-uh-CRON</i>
π, Π	pi <i>PIE</i>
ρ	rho <i>ROW</i>
σ, Σ	sigma <i>SIG-muh</i>
τ	tau <i>TOW (as in cow)</i>
υ, Υ	upsilon <i>OOP-suh-LON</i>
ϕ, Φ	phi <i>FEE, or FI (as in hi)</i>
χ	chi <i>KI (as in hi)</i>
ψ, Ψ	psi <i>SIGH, or PSIGH</i>
ω, Ω	omega <i>oh-MAY-guh</i>

Appendix B

Solutions to Exercises

B.1 Solutions to Warm-up Questions

Linear Algebra

Vectors

Define the vectors $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$, $\mathbf{v} = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}$, and the scalar $c = 2$.

1. $\mathbf{u} + \mathbf{v} = \begin{pmatrix} 5 \\ 7 \\ 9 \end{pmatrix}$

2. $c\mathbf{v} = \begin{pmatrix} 8 \\ 10 \\ 12 \end{pmatrix}$

3. $\mathbf{u} \cdot \mathbf{v} = 1(4) + 2(5) + 3(6) = 32$

If you are having trouble with these problems, please review Section 2.1.

Are the following sets of vectors linearly independent?

1. $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $\mathbf{v} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$

\rightsquigarrow No:

$$2\mathbf{u} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \mathbf{v} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$

so infinitely many linear combinations of \mathbf{u} and \mathbf{v} that amount to $\mathbf{0}$ exist.

$$2. \mathbf{u} = \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix}, \mathbf{v} = \begin{pmatrix} 3 \\ 7 \\ 9 \end{pmatrix}$$

↪ Yes: we cannot find linear combination of these two vectors that would amount to zero.

$$3. \mathbf{a} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 3 \\ -4 \\ -2 \end{pmatrix}, \mathbf{c} = \begin{pmatrix} 5 \\ -10 \\ -8 \end{pmatrix}$$

↪ No: After playing around with some numbers, we can find that

$$-2\mathbf{a} = \begin{pmatrix} -4 \\ 2 \\ -2 \end{pmatrix}, 3\mathbf{b} = \begin{pmatrix} 9 \\ -12 \\ -6 \end{pmatrix}, -1\mathbf{c} = \begin{pmatrix} -5 \\ 10 \\ 8 \end{pmatrix}$$

So

$$-2\mathbf{a} + 3\mathbf{b} - \mathbf{c} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

i.e., a linear combination of these three vectors that would amount to zero exists. If you are having trouble with these problems, please review Section 2.2.

Matrices

$$\mathbf{A} = \begin{pmatrix} 7 & 5 & 1 \\ 11 & 9 & 3 \\ 2 & 14 & 21 \\ 4 & 1 & 5 \end{pmatrix}$$

What is the dimensionality of matrix \mathbf{A} ? 4×3 What is the element a_{23} of \mathbf{A} ? 3 Given that

$$\mathbf{B} = \begin{pmatrix} 1 & 2 & 8 \\ 3 & 9 & 11 \\ 4 & 7 & 5 \\ 5 & 1 & 9 \end{pmatrix}$$

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} 8 & 7 & 9 \\ 14 & 18 & 14 \\ 6 & 21 & 26 \\ 9 & 2 & 14 \end{pmatrix}$$

Given that

$$\mathbf{C} = \begin{pmatrix} 1 & 2 & 8 \\ 3 & 9 & 11 \\ 4 & 7 & 5 \end{pmatrix}$$

$\mathbf{A} + \mathbf{C}$ = No solution, matrices non-conformable

Given that

$$c\mathbf{A} = \begin{matrix} c = 2 \\ \begin{pmatrix} 14 & 10 & 2 \\ 22 & 18 & 6 \\ 4 & 28 & 42 \\ 8 & 2 & 10 \end{pmatrix} \end{matrix}$$

If you are having trouble with these problems, please review Section 2.3.

Operations

Summation

Simplify the following

1. $\sum_{i=1}^3 i = 1 + 2 + 3 = 6$
2. $\sum_{k=1}^3 (3k + 2) = 3 \sum_{k=1}^3 k + \sum_{k=1}^3 2 = 3 \times 6 + 3 \times 2 = 24$
3. $\sum_{i=1}^4 (3k + i + 2) = 3 \sum_{i=1}^4 k + \sum_{i=1}^4 i + \sum_{i=1}^4 2 = 12k + 10 + 8 = 12k + 18$

Products

1. $\prod_{i=1}^3 i = 1 \cdot 2 \cdot 3 = 6$
2. $\prod_{k=1}^3 (3k + 2) = (3 + 2) \cdot (6 + 2) \cdot (9 + 2) = 440$

To review this material, please see Section 3.1.

Logs and exponents

Simplify the following

1. $4^2 = 16$
2. $4^2 2^3 = 2^{2 \cdot 2} 2^3 = 2^{4+3} = 128$
3. $\log_{10} 100 = \log_{10} 10^2 = 2$
4. $\log_2 4 = \log_2 2^2 = 2$
5. when \log is the natural log, $\log e = \log_e e^1 = 1$

6. when a, b, c are each constants, $e^a e^b e^c = e^{a+b+c}$,
7. $\log 0 = \text{undefined}$ – no exponentiation of anything will result in a 0.
8. $e^0 = 1$ – any number raised to the 0 is always 1.
9. $e^1 = e$ – any number raised to the 1 is always itself
10. $\log e^2 = \log_e e^2 = 2$

To review this material, please see Section 3.3

Limits

Find the limit of the following.

1. $\lim_{x \rightarrow 2} (x - 1) = 1$
2. $\lim_{x \rightarrow 2} \frac{(x-2)(x-1)}{(x-2)} = 1$, though note that the original function $\frac{(x-2)(x-1)}{(x-2)}$ would have been undefined at $x = 2$ because of a divide by zero problem; otherwise it would have been equal to $x - 1$.
3. $\lim_{x \rightarrow 2} \frac{x^2-3x+2}{x-2} = 1$, same as above.

To review this material please see Section 4.5

Calculus

For each of the following functions $f(x)$, find the derivative $f'(x)$ or $\frac{d}{dx} f(x)$

1. $f(x) = c, f'(x) = 0$
2. $f(x) = x, f'(x) = 1$
3. $f(x) = x^2, f'(x) = 2x$
4. $f(x) = x^3, f'(x) = 3x^2$
5. $f(x) = 3x^2 + 2x^{1/3}, f'(x) = 6x + \frac{2}{3}x^{-2/3}$
6. $f(x) = (x^3)(2x^4), f'(x) = \frac{d}{dx} 2x^7 = 14x^6$

For a review, please see Section 5.2 - 5.3

Optimization

For each of the following functions $f(x)$, does a maximum and minimum exist in the domain $x \in \mathbf{R}$? If so, for what are those values and for which values of x ?

1. $f(x) = x \rightsquigarrow$ neither exists.
2. $f(x) = x^2 \rightsquigarrow$ a minimum $f(x) = 0$ exists at $x = 0$, but not a maximum.
3. $f(x) = -(x-2)^2 \rightsquigarrow$ a maximum $f(x) = 0$ exists at $x = 2$, but not a minimum.

If you are stuck, please try sketching out a picture of each of the functions.

Probability

1. If there are 12 cards, numbered 1 to 12, and 4 cards are chosen, $\binom{12}{4} = \frac{12 \cdot 11 \cdot 10 \cdot 9}{4!} = 495$ possible hands exist (unordered, without replacement) .
2. Let $A = \{1, 3, 5, 7, 8\}$ and $B = \{2, 4, 7, 8, 12, 13\}$. Then $A \cup B = \{1, 2, 3, 4, 5, 7, 8, 12, 13\}$, $A \cap B = \{7, 8\}$? If A is a subset of the Sample Space $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, then the complement $A^C = \{2, 4, 6, 9, 10\}$
3. If we roll two fair dice, what is the probability that their sum would be 11? $\rightsquigarrow \frac{1}{18}$
4. If we roll two fair dice, what is the probability that their sum would be 12? $\rightsquigarrow \frac{1}{36}$. There are two independent dice, so $6^2 = 36$ options in total. While the previous question had two possibilities for a sum of 11 (5,6 and 6,5), there is only one possibility out of 36 for a sum of 12 (6,6).

B.2 Solutions to Linear Algebra Exercises

Example 2.1:

1. $\begin{pmatrix} -1 & -3 & -3 \end{pmatrix}$
2. $6 + 4 + 10 = 20$

Exercise 2.1:

1. $\begin{pmatrix} -2 & 4 & -7 & -5 \end{pmatrix}$
2. $\begin{pmatrix} 2 & 26 & -14 & 4 & 30 \end{pmatrix}$
3. $63 - 3 - 10 + 24 = 74$
4. undefined

Example 2.2:

1. yes
2. no

Exercise 2.2:

1. yes
2. no ($-v_1 - v_2 + v_3 = 0$)

Example 2.3:

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} 2 & 4 & 4 \\ 6 & 6 & 8 \end{pmatrix}$$

Example 2.4:

$$s\mathbf{A} = \begin{pmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \end{pmatrix}$$

Example 2.5:

1. $\begin{pmatrix} aA + bC & aB + bD \\ cA + dC & cB + dD \\ eA + fC & eB + fD \end{pmatrix}$
2. $\begin{pmatrix} 1(-2) + 2(4) - 1(2) & 1(5) + 2(-3) - 1(1) \\ 3(-2) + 1(4) + 4(2) & 3(5) + 1(-3) + 4(1) \end{pmatrix} = \begin{pmatrix} 4 & -2 \\ 6 & 16 \end{pmatrix}$

Exercise 2.3:

1. $AB = \begin{pmatrix} 4 & 11 & -15 \\ 5 & 7 & -7 \end{pmatrix}$
2. $BA = \text{undefined}$
3. $(BC)^T = \text{undefined}$
4. $BC^T = \begin{pmatrix} 1 & 5 & -7 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \\ 2 & 0 & 0 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 2 & 4 \\ -1 & 6 \end{pmatrix} = \begin{pmatrix} 20 & -22 \\ 5 & 4 \\ -3 & 2 \\ 6 & 0 \end{pmatrix}$

Exercise 2.4:

There are many answers to this. Some possible simple ones are as follows:

1. One solution:

$$\begin{array}{rcl} -x & + & y = 0 \\ x & + & y = 2 \end{array}$$

2. No solution:

$$\begin{array}{rcl} -x & + & y = 0 \\ x & - & y = 2 \end{array}$$

3. Infinite solutions:

$$\begin{array}{rcl} -x & + & y = 0 \\ 2x & - & 2y = 0 \end{array}$$

Exercise 2.5:

$$\left(\begin{array}{cccc|cc} 2 & -7 & 9 & -4 & 0 & 0 & 8 \\ 0 & 41 & 9 & 0 & 0 & 5 & 11 \\ 1 & -15 & 0 & 0 & -11 & 0 & 9 \end{array} \right)$$

Example 2.9:

$$\begin{array}{rcl} x & - & 3y = -3 \\ 2x & + & y = 8 \end{array}$$

$$\begin{array}{rcl} x & - & 3y = -3 \\ & & 7y = 14 \end{array}$$

$$\begin{array}{rcl} x & - & 3y = -3 \\ & & y = 2 \end{array}$$

$$\begin{array}{rcl} x & = & 3 \\ y & = & 2 \end{array}$$

Exercise 2.6:

1. $x = 2, y = 2, z = -1$
2. $x = -17, y = -3, z = -35$

Exercise 2.7:

1. rank is 2
2. rank is 3

Exercise 2.8:

$$\mathbf{A}^{-1} = \begin{pmatrix} 1 & 0 & -4 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Exercise 2.9:

$$\mathbf{z} = \mathbf{A}^{-1}\mathbf{b} = \begin{pmatrix} 1/5 & 4/5 \\ 2/5 & 3/5 \end{pmatrix} \begin{pmatrix} 5 \\ -10 \end{pmatrix} = \begin{pmatrix} -7 \\ -4 \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}$$

Exercise 2.10:

1. nonsingular
2. singular

Exercise 2.11:

$$\begin{pmatrix} \frac{2}{41} & \frac{-5}{41} \\ \frac{7}{41} & \frac{3}{41} \end{pmatrix}$$

B.3 Solutions to Functions and Operations Exercises

Example 3.1:

1. $1 + 2 + 3 + 4 + 5 = 15$
2. $1 \times 2 \times 3 \times 4 \times 5 = 120$
3. 2
4. $4 \cdot (4 - 1) \cdot (4 - 2) \cdot (4 - 3) = 4 \times 3 \times 2 \times 1 = 24$

Exercise 3.1:

1. $7(4 + 3 + 7) = 98$
2. $2 + 2 + 2 + 2 + 2 = 10$
3. $2^3(7)(11)(2) = 1232$

Example 3.2:

1. one-to-one
2. many-to-one

Exercise 3.2:

1. many-to-one
2. one-to-one

Example 3.3:

1. 2
2. 4
3. $3 \log_4(x) + 5 \log_4(y)$

Exercise 3.3:

1. 3
2. $9 \log(x) + 5 \log(y) - 3 \log(z)$
3. $\frac{1}{2}(\ln x + \ln y)$

Example 3.4:

1. $y = 3x + 2 \implies -3x = 2 - y \implies 3x = y - 2 \implies x = \frac{1}{3}(y - 2)$
2. $x = \ln y$

Exercise 3.4:

1. $\frac{-2}{3}$
2. $x = \{1, -4\}$
3. $x = -\ln 10$

B.4 Solutions to Limits Exercises

Example 4.1:

1. $\{A_n\} = \left\{2 - \frac{1}{n^2}\right\} = \left\{1, \frac{7}{4}, \frac{17}{9}, \frac{31}{16}, \frac{49}{25}, \dots\right\} = 2$
2. $\{B_n\} = \left\{\frac{n^2+1}{n}\right\} = \left\{2, \frac{5}{2}, \frac{10}{3}, \frac{17}{4}, \dots\right\}$

$$3. \{C_n\} = \{(-1)^n (1 - \frac{1}{n})\} = \{0, \frac{1}{2}, -\frac{2}{3}, \frac{3}{4}, -\frac{4}{5}\}$$

Exercise 4.3:

(See chapter)

Example 4.3:

$$1. k$$

$$2. c$$

$$3. \lim_{x \rightarrow 2} (2x - 3) = 2 \lim_{x \rightarrow 2} x - 3 \lim_{x \rightarrow 2} 1 = 1$$

$$4. \lim_{x \rightarrow c} x^n = \lim_{x \rightarrow c} x \cdots [\lim_{x \rightarrow c} x] = c \cdots c = c^n$$

Exercise 4.4:

Although this function seems large, the thing our eyes should focus on is where the highest order polynomial remains. That will grow the fastest, so if the highest order term is on the denominator, the fraction will converge to 0, if it is on the numerator it will converge to negative infinity. Previewing the multiplication by hand, we can see that the $-x^9$ on the numerator will be the largest power. So the answer will be $-\infty$. We can also confirm this by writing out fractions:

$$\begin{aligned} & \lim_{x \rightarrow \infty} \frac{(1 + \frac{3}{x^3} - \frac{99}{4x^4}) (-\frac{2}{x^5} + 1)}{(1 + \frac{9}{18x} - \frac{3}{18x^5} - \frac{1}{18x^7}) (1 + \frac{1}{x})} \\ & \times \frac{x^4}{1} \times -\frac{x^5}{1} \times \frac{1}{18x^7} \times \frac{1}{x} \\ & = 1 \times \lim_{-x \rightarrow \infty} \frac{x}{18} \end{aligned}$$

Exercise 4.6: See Figure B.1. Divide each part by x , and we get $x + \frac{2}{x}$ on the numerator, 1 on the denominator. So, without worrying about a function being not defined, we can say $\lim_{x \rightarrow 0} f(x) = 0$.

B.5 Solutions to Calculus Exercises

Exercise 5.1:

$$1. f'(x) = 0$$

$$2. f'(x) = 1$$

$$3. f'(x) = 2x^3$$

$$4. f'(x) = 3x^2$$

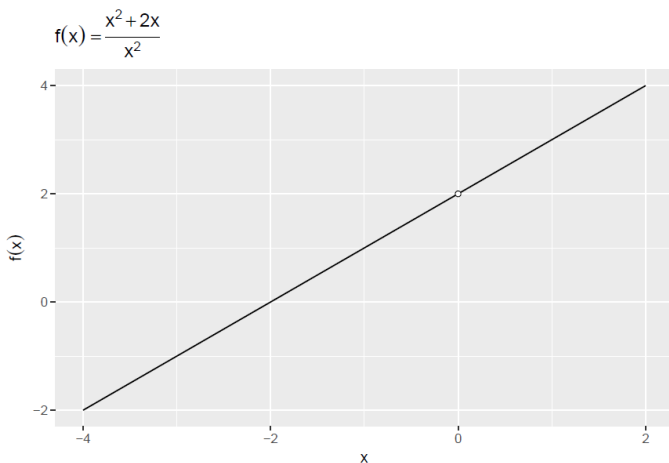


Figure B.1: A function undefined at $x = 0$

5. $f(x) = -2x^{-3}$

6. $f(x) = 14x^6$

7. $f(x) = 4x^3 - 3x^2 + 2x - 1$

8. $f(x) = 5x^4 + 3x^2 - 2x$

9. $f(x) = 6x + \frac{2}{3}x^{-\frac{2}{3}}$

10. $f(x) = \frac{-4x}{x^4 - 2x^2 + 1}$

Example 5.3:

For convenience, define $f(z) = z^6$ and $z = g(x) = 3x^2 + 5x - 7$. Then, $y = f[g(x)]$ and

$$\begin{aligned} \frac{d}{dx}y &= f'(z)g'(x) \\ &= 6(3x^2 + 5x - 7)^5(6x + 5) \end{aligned}$$

Example 5.4:

1. Let $u(x) = -3x$. Then $u'(x) = -3$ and $f'(x) = -3e^{-3x}$.

2. Let $u(x) = x^2$. Then $u'(x) = 2x$ and $f'(x) = 2xe^{x^2}$.

Example 5.5:

1. Let $u(x) = x^2 + 9$. Then $u'(x) = 2x$ and

$$\frac{dy}{dx} = \frac{u'(x)}{u(x)} = \frac{2x}{(x^2 + 9)}$$

2. Let $u(x) = \log x$. Then $u'(x) = 1/x$ and $\frac{dy}{dx} = \frac{1}{(x \log x)}$.

3. Use the generalized power rule.

$$\frac{dy}{dx} = \frac{(2 \log x)}{x}$$

4. We know that $\log e^x = x$ and that $dx/dx = 1$, but we can double check. Let $u(x) = e^x$. Then $u'(x) = e^x$ and $\frac{dy}{dx} = \frac{u'(x)}{u(x)} = \frac{e^x}{e^x} = 1$.

Example 5.9:

What is $F(x)$? From the power rule, recognize $\frac{d}{dx}x^3 = 3x^2$ so

$$\begin{aligned} F(x) &= x^3 \\ \int_1^3 f(x)dx &= F(x=3) - F(x=1) \\ &= 3^3 - 1^3 \\ &= 26 \end{aligned}$$

Example 5.10:

The problem here is the $\sqrt{x+1}$ term. However, if the integrand had \sqrt{x} times some polynomial, then we'd be in business. Let's try $u = x + 1$. Then $x = u - 1$ and $dx = du$. Substituting these into the above equation, we get

$$\begin{aligned} \int x^2 \sqrt{x+1} dx &= \int (u-1)^2 \sqrt{u} du \\ &= \int (u^2 - 2u + 1) u^{1/2} du \\ &= \int (u^{5/2} - 2u^{3/2} + u^{1/2}) du \end{aligned}$$

We can easily integrate this, since it is just a polynomial. Doing so and substituting $u = x + 1$ back in, we get

$$\int x^2 \sqrt{x+1} dx = 2(x+1)^{3/2} \left[\frac{1}{7}(x+1)^2 - \frac{2}{5}(x+1) + \frac{1}{3} \right] + c$$

Example 5.11:

When an expression is raised to a power, it is often helpful to use this

expression as the basis for a substitution. So, let $u = 1 + e^{2x}$. Then $du = 2e^{2x}dx$ and we can set $5e^{2x}dx = 5du/2$. Additionally, $u = 2$ when $x = 0$ and $u = 1 + e^2$ when $x = 1$. Substituting all of this in, we get

$$\begin{aligned}\int_0^1 \frac{5e^{2x}}{(1 + e^{2x})^{1/3}} dx &= \frac{5}{2} \int_2^{1+e^2} \frac{du}{u^{1/3}} \\ &= \frac{5}{2} \int_2^{1+e^2} u^{-1/3} du \\ &= \frac{15}{4} u^{2/3} \Big|_2^{1+e^2} \\ &= 9.53\end{aligned}$$

Exercise 5.9:

1.

$$\int x^n e^{ax} dx$$

As in the first problem, let

$$u = x^n, dv = e^{ax} dx$$

Then $du = nx^{n-1}dx$ and $v = (1/a)e^{ax}$.

Substituting these into the integration by parts formula gives

$$\begin{aligned}\int x^n e^{ax} dx &= uv - \int v du \\ &= x^n \left(\frac{1}{a} e^{ax} \right) - \int \frac{1}{a} e^{ax} nx^{n-1} dx \\ &= \frac{1}{a} x^n e^{ax} - \frac{n}{a} \int x^{n-1} e^{ax} dx\end{aligned}$$

Notice that we now have an integral similar to the previous one, but with x^{n-1} instead of x^n .

For a given n , we would repeat the integration by parts procedure until the integrand was directly integratable — e.g., when the integral became $\int e^{ax} dx$.

2.

$$\int x^3 e^{-x^2} dx$$

We could, as before, choose $u = x^3$ and $dv = e^{-x^2} dx$. But we can't then find v — i.e., integrating $e^{-x^2} dx$ isn't possible. Instead, notice that

$$\frac{d}{dx} e^{-x^2} = -2xe^{-x^2},$$

which can be factored out of the original integrand

$$\int x^3 e^{-x^2} dx = \int x^2 (x e^{-x^2}) dx.$$

We can then let $u = x^2$ and $dv = x e^{-x^2} dx$. Then $du = 2x dx$ and $v = -\frac{1}{2} e^{-x^2}$. Substituting these in, we have

$$\begin{aligned} \int x^3 e^{-x^2} dx &= uv - \int v du \\ &= x^2 \left(-\frac{1}{2} e^{-x^2} \right) - \int \left(-\frac{1}{2} e^{-x^2} \right) 2x dx \\ &= -\frac{1}{2} x^2 e^{-x^2} + \int x e^{-x^2} dx \\ &= -\frac{1}{2} x^2 e^{-x^2} - \frac{1}{2} e^{-x^2} + c \end{aligned}$$

B.6 Solutions to Optimization Exercises

Exercise 6.1:

Solve using the following steps:

1. Rewrite with the slack variables:

$$\begin{aligned} \max_{x_1, x_2} f(x) &= -(x_1^2 + 2x_2^2) \text{ s.t. } \begin{aligned} x_1 + x_2 &\leq 4 - s_1^2 \\ -x_1 &\leq 0 - s_2^2 \\ -x_2 &\leq 0 - s_3^2 \end{aligned} \end{aligned}$$

2. Write the Lagrangian:

$$\begin{aligned} L(x_1, x_2, \lambda_1, \lambda_2, \lambda_3, s_1, s_2, s_3) \\ &= -(x_1^2 + 2x_2^2) - \lambda_1(x_1 + x_2 + s_1^2 - 4) \\ &\quad - \lambda_2(-x_1 + s_2^2) - \lambda_3(-x_2 + s_3^2) \end{aligned}$$

3. Take the partial derivatives and set equal to zero:

$$\begin{aligned} \frac{\partial L}{\partial x_1} &= -2x_1 - \lambda_1 + \lambda_2 = 0 \\ \frac{\partial L}{\partial x_2} &= -4x_2 - \lambda_1 + \lambda_3 = 0 \\ \frac{\partial L}{\partial \lambda_1} &= -(x_1 + x_2 + s_1^2 - 4) = 0 \\ \frac{\partial L}{\partial \lambda_2} &= -(-x_1 + s_2^2) = 0 \\ \frac{\partial L}{\partial \lambda_3} &= -(-x_2 + s_3^2) = 0 \\ \frac{\partial L}{\partial s_1} &= 2s_1 \lambda_1 = 0 \\ \frac{\partial L}{\partial s_2} &= 2s_2 \lambda_2 = 0 \\ \frac{\partial L}{\partial s_3} &= 2s_3 \lambda_3 = 0 \end{aligned}$$

4. Consider all ways that the complementary slackness conditions are solved:

Hypothesis	s_1	s_2	s_3	λ_1	λ_2	λ_3	x_1	x_2	$f(x_1, x_2)$
$s_1 = s_2 = s_3 = 0$	No solution								
$s_1 \neq 0, s_2 = s_3 = 0$	2	0	0	0	0	0	0	0	0
$s_2 \neq 0, s_1 = s_3 = 0$	0	2	0	-8	0	-8	4	0	-16
$s_3 \neq 0, s_1 = s_2 = 0$	0	0	2	-16	-16	0	0	4	-32
$s_1 \neq 0, s_2 \neq 0, s_3 = 0$	No solution								
$s_1 \neq 0, s_3 \neq 0, s_2 = 0$	No solution								
$s_2 \neq 0, s_3 \neq 0, s_1 = 0$	0	$\sqrt{\frac{8}{3}}$	$\sqrt{\frac{4}{3}}$	$-\frac{16}{3}$	0	0	$\frac{8}{3}$	$\frac{4}{3}$	$-\frac{32}{3}$
$s_1 \neq 0, s_2 \neq 0, s_3 \neq 0$	No solution								

This shows that there are four critical points: $(0, 0)$, $(4, 0)$, $(0, 4)$, and $(\frac{8}{3}, \frac{4}{3})$

5. Find maximum: Looking at the values of $f(x_1, x_2)$ at the critical points, we see that the constrained maximum is located at $(x_1, x_2) = (0, 0)$, which is the same as the unconstrained max. The constrained minimum is located at $(x_1, x_2) = (0, 4)$, while there is no unconstrained minimum for this problem.

Exercise 6.2:

Solve using the following steps:

1. Write the Lagrangian:

$$L(x_1, x_2, \lambda) = \frac{1}{3} \log(x_1 + 1) + \frac{2}{3} \log(x_2 + 1) - \lambda(x_1 + 2x_2 - 4)$$

2. Find the First Order Conditions:

Kuhn-Tucker Conditions

$$\frac{\partial L}{\partial x_1} = \frac{1}{3(x_1+1)} - \lambda \leq 0$$

$$\frac{\partial L}{\partial x_2} = \frac{2}{3(x_2+1)} - \lambda \leq 0$$

$$\frac{\partial L}{\partial \lambda} = -(x_1 + 2x_2 - 4) \geq 0$$

Complementary Slackness Conditions

$$x_1 \frac{\partial L}{\partial x_2} = x_1 \left(\frac{1}{3(x_2+1)} - \lambda \right) = 0$$

$$x_2 \frac{\partial L}{\partial x_2} = x_2 \left(\frac{2}{3(x_2+1)} - \lambda \right) = 0$$

$$\lambda \frac{\partial L}{\partial \lambda} = -\lambda(x_1 + 2x_2 - 4) = 0$$

Non-negativity Conditions

$$x_1 \geq 0$$

$$x_2 \geq 0$$

$$\lambda \geq 0$$

3. Consider all border and interior cases:

Hypothesis	λ	x_1	x_2	$f(x_1, x_2)$
$x_1 = 0, x_2 = 0$	No Solution			
$x_1 = 0, x_2 \neq 0$	No Solution			
$x_1 \neq 0, x_2 = 0$	No Solution			
$x_1 \neq 0, x_2 \neq 0$		$\frac{4}{3}$	$\frac{4}{3}$	$\log \frac{7}{3}$

4. Find Maximum:

Three of the critical points violate the constraints, so the point $(\frac{4}{3}, \frac{4}{3})$ is the maximum.

B.7 Solutions to Probability Theory Exercises

Example 7.2:

- $5 \times 5 \times 5 = 125$
- $5 \times 4 \times 3 = 60$
- $\binom{5}{3} = \frac{5!}{(5-3)!3!} = \frac{5 \times 4}{2 \times 1} = 10$

Exercise 7.1:

- $\binom{52}{4} = \frac{52!}{(52-4)!4!} = 270725$

Example 7.3:

- $\{1, 2, 3, 4, 5, 6\}$
- $\{5, 6\}$
- $\{1, 2, 7, 8, 9, 10\}$
- $\{3, 4\}$

Exercise 7.2:

Sample Space: $\{2, 3, 4, 5, 6, 7, 8\}$

- $\{3, 4, 5, 6, 7\}$
- $\{4, 5, 6\}$

Example 7.4:

1. 1, 2, 3, 4, 5, 6
2. $\frac{1}{6}$
3. 0
4. $\frac{1}{2}$
5. $\frac{4}{6} = \frac{2}{3}$
6. 1
7. $A \cup B = \{1, 2, 3, 4, 6\}$, $A \cap B = \{2\}$, $\frac{5}{6}$

Exercise 7.3:

1. $P(X = 5) = \frac{4}{16}$, $P(X = 3) = \frac{2}{16}$, $P(X = 6) = \frac{3}{16}$
2. What is $P(X = 5 \cup X = 3)^C = \frac{10}{16}$?

Example 7.5:

1. $\frac{n_{ab} + n_{abc}}{N}$
2. $\frac{n_{ab} + n_{acb}}{N}$
3. $\frac{n_{ab}}{N}$
4. $\frac{\frac{n_{ab}}{N}}{\frac{n_{ab} + n_{acb}}{N}} = \frac{n_{ab}}{n_{ab} + n_{acb}}$
5. $\frac{\frac{n_{ab}}{N}}{\frac{n_{ab} + n_{abc}}{N}} = \frac{n_{ab}}{n_{ab} + n_{abc}}$

Example 7.6:

$$P(1|Odd) = \frac{P(1 \cap Odd)}{P(Odd)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

Example 7.7:

Using Bayes' Law and the Law of Total Probability, we know: Define A as a positive test. Define B_1 as having cancer. Define B_2 as not having cancer. We want to know $P(B_1|A)$.

$$\begin{aligned} P(B_1|A) &= \frac{P(B_1)P(A|B_1)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_1)} \\ &= \frac{.1 \times .9}{.1 \times .9 + .9 \times .1} \\ &= .5 \end{aligned}$$

Exercise 7.4:

We are given that

$$P(G) = .02, P(C|G) = .95, P(C^c|G^c) = .97$$

$$\begin{aligned} P(G|C) &= \frac{P(C|G)P(G)}{P(C)} \\ &= \frac{P(C|G)P(G)}{P(C|G)P(G) + P(C|G^c)P(G^c)} \\ &= \frac{P(C|G)P(G)}{P(C|G)P(G) + [1 - P(C^c|G^c)]P(G^c)} \\ &= \frac{.95 \times .02}{.95 \times .02 + .03 \times .98} \\ &= .38 \end{aligned}$$

Example 7.11:

$$E(Y) = 7/2$$

We would never expect the result of a rolled die to be $7/2$, but that would be the average over a large number of rolls of the die.

Example 7.12:

$$0.75$$

Example 7.13:

$$E(X) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{3}{2}$$

Since there is a 1-to-1 mapping from X to X^2 :

$$E(X^2) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 4 \times \frac{3}{8} + 9 \times \frac{1}{8} = \frac{24}{8} = 3$$

$$\begin{aligned} \text{Var}(x) &= E(X^2) - E(x)^2 \\ &= 3 - \left(\frac{3}{2}\right)^2 \\ &= \frac{3}{4} \end{aligned}$$

Exercise 7.6:

$$1. E(X) = -2\left(\frac{1}{5}\right) + -1\left(\frac{1}{6}\right) + 0\left(\frac{1}{5}\right) + 1\left(\frac{1}{15}\right) + 2\left(\frac{11}{30}\right) = \frac{7}{30}$$

$$2. E(Y) = 0(\frac{1}{5}) + 1(\frac{7}{30}) + 4(\frac{17}{30}) = \frac{5}{2}$$

$$3. \text{Var}(X) = E[X^2] - E[X]^2 = E(Y) - E(X)^2 = \frac{5}{2} - \frac{7}{30}^2 \approx 2.45$$

Exercise 7.7:

expectation = $\frac{6}{5}$, variance = $\frac{6}{25}$

Exercise 7.8:

$$1. \text{ mean} = 2, \text{ standard deviation} = \sqrt{(\frac{2}{3})}$$

$$2. \frac{1}{8}(2 - \sqrt{(\frac{2}{3})})^2$$

Index

- Derivative,
 - Properties, 49,
 - Partial Derivatives, 55,
- Exponent, 30
- Integral
 - Definite, 59
 - Properties, 58
 - Indefinite, 56
 - Substitution, 62,
 - Integration By Parts, 63,
- Limits,
 - Properties, 42,
 - Sequence, 40,
 - Functions, 41,
- Log, 30
- Matrices, 8
 - Determinant, 21
 - Inverse, 18
 - Linear Independence, 7
 - Rank, 17
- Operators,
 - Factorial, 26
 - Modulo, 26
 - Product, 26
 - Sum, 25
- Systems of Equations, 13, 15, 21,
- Taylor Series, 56,
- Vectors, 5