Quantitative Finance Handbook

Xi Tan (xtan3.1415926@gmail.com)

February 7, 2019

Contents

Preface

This book project, which consists of four subjects: Finance, Mathematics, Statistics, and Computer Science, is tailored specifically to prepare someone for a quant career. It originated from my general belief of the hierarchy of solving a problem — problems are solved at strategic, tactical, and operational levels.

Microeconomics and *Macroeconomics* explain the driving forces of capital markets, from a legislator's perspective. *Accounting* and *Corporate Finance* take a closer and necessary look at these forces, from a different angle. *Stochastic Calculus* and *Asset Pricing* provide with a set of tools and ideas that enables us to **strategically** model one of the central problems in Quantitative Finance.

Generally speaking, there are two paths to solve a quantitative finance problem at the **tactical** level: the mathematical way and the statistical way. There are only two pieces of math we need to know: *Analysis*, in particular measure-theoretical probability and differential equations; and *Linear Algebra*, with functional analysis in mind. Statistics, on the other hand, should start with *Statistical Experiment Design*, from which we learn how to collect data for statistical models. Next, the study of *Random Variables* and *Stochastic Processes* introduce the building blocks of the statistical "pillbox", with *Mathematical Statistics* the "scaffold". Once the "pillbox" is ready, we are equipped to tackle our problems using *Machine Learning*, which is essentially a collection of statistical models and optimization algorithms.

Computer Architecture and *Operating System* are respectively about the "hardware" and "software" of a single computer. The interaction of multiple computers is understood in *Computer Network*. Once we are comfortable with these concepts, we will be able to use *Data Structure and Algorithms* to solve problems at the **operational** level, and use *C++* and/or *Java* to implement our ideas.

I am aware that it can take a while, and even multiple advanced degrees, to finish this curriculum, but let's remember the motto from the Leipzig Gewandhaus Orchestra: "***Res severa est verum gaudium***".

Xi Tan
West Lafayette, IN
October, 2013

Part I

Finance and Economics

Chapter 1

Economical Finance

1.1 Capital Market Overview

1.2 Trading and Exchanges

1.3 Macroeconomic Environment

1.4 Classic Finance Theory

1.4.1 Time Value of Money

1.4.2 Capital Asset Pricing Model (CAPM)

Chapter 2

Mathematical Finance

2.1 Probability Theory

2.1.1 Probability Space

2.1.2 Information and σ-Algebra

2.1.3 Conditional Expectation

2.1.4 Martingales

2.1.5 Change of Measure and Girsanov's Theorem

2.2 Brownian Motion and Stochastic Calculus

2.2.1 Stochastic Processes

2.2.2 Brownian Motion

2.2.3 Geometric Brownian Motion

2.2.4 Itô Integral and Itô Process

2.2.5 Function of Itô Processes and Itô's Lemma

2.3 Stochastic Differential Equations

2.3.1 The Feynman–Kac Formula

2.3.2 Kolmogorov Forward and Backward Equations

Start with the SDE defined by

$$\mathrm{d}\,X_t = \mu(X_t)\,\mathrm{d}\,t + \sigma(X_t)\,\mathrm{d}\,W_t. \tag{2.1}$$

The transition density $\rho(x,t|y,s)$ is defined by

$$\int_A \rho(x,t|y,s)\,\mathrm{d}\,x = \mathbb{P}[X_{t+s} \in A|X_s = y] = \mathbb{P}[X_t \in A|X_0 = y]. \tag{2.2}$$

The density $\rho(x,t|y,s)$ is time-invariance since $\mu(X_t)$ and $\sigma(X_t)$ are assumed to be time invariance, and consequently, that X_t is assumed to be stationary.

Consider a differentiable function $V(X_t,t) = V(x,t)$ with $V(X_t,t) = 0$ for $t \notin (0,T)$. Then by Itô's lemma

$$\mathrm{d}\,V = \left[\frac{\partial V}{\partial t} + \mu\frac{\partial V}{\partial x} + \frac{1}{2}\sigma^2\frac{\partial^2 V}{\partial x^2}\right]\mathrm{d}\,t + \left[\sigma\frac{\partial V}{\partial x}\right]\mathrm{d}\,W_t \tag{2.3}$$

so that

$$V(X_T,T) - V(X_0,0) = \int_0^T \left[\frac{\partial V}{\partial t} + \mu\frac{\partial V}{\partial x} + \frac{1}{2}\sigma^2\frac{\partial^2 V}{\partial x^2}\right]\mathrm{d}\,t + \int_0^T \left[\sigma\frac{\partial V}{\partial x}\right]\mathrm{d}\,W_t \tag{2.4}$$

where $\mu = \mu(X_t)$ and $\sigma = \sigma(X_t)$ for notational convenience. Take the conditional expectation of both sides of the above equation given X_0

$$\begin{aligned}\mathbb{E}[V(X_T,T) - V(X_0,0)] &= \mathbb{E}\left[\int_0^T \left[\frac{\partial V}{\partial t} + \mu\frac{\partial V}{\partial x} + \frac{1}{2}\sigma^2\frac{\partial^2 V}{\partial x^2}\right]\mathrm{d}\,t\right] \\ &= \int_{\mathbb{R}}\left\{\int_0^T \left[\frac{\partial V}{\partial t} + \mu\frac{\partial V}{\partial x} + \frac{1}{2}\sigma^2\frac{\partial^2 V}{\partial x^2}\right]\mathrm{d}\,t\right\}\rho(x,t|y,s)\,\mathrm{d}\,x\end{aligned} \tag{2.5}$$

We write this equation as the sum of three integrals:

$$I_1 \equiv \int_{\mathbb{R}}\int_0^T \rho\frac{\partial V}{\partial t}\,\mathrm{d}\,t\,\mathrm{d}\,x \tag{2.6}$$

$$I_2 \equiv \int_{\mathbb{R}}\int_0^T \rho\mu\frac{\partial V}{\partial x}\,\mathrm{d}\,t\,\mathrm{d}\,x \tag{2.7}$$

$$I_3 \equiv \frac{1}{2}\int_{\mathbb{R}}\int_0^T \rho\sigma^2\frac{\partial^2 V}{\partial x^2}\,\mathrm{d}\,t\,\mathrm{d}\,x \tag{2.8}$$

The objective of the derivation is to apply integration by parts to get rid of the derivatives of V.

Evaluation of the Integrals

The trick is that I_1 is evaluated using integration by parts on t, while I_2 and I_3 are each evaluated using integration by parts on x.

Evaluation of I_1

$$I_1 = \int_{\mathbb{R}} \left[\int_0^T \rho \frac{\partial V}{\partial t} \,\mathrm{d}\, t \right] \mathrm{d}\, x = \int_{\mathbb{R}} \left[\rho V|_0^T - \int_0^T \frac{\partial \rho}{\partial t} V \,\mathrm{d}\, t \right] \mathrm{d}\, x = -\int_{\mathbb{R}} \int_0^T \frac{\partial \rho}{\partial t} V \,\mathrm{d}\, t \,\mathrm{d}\, x \tag{2.9}$$

where we have used the fact that at boundaries 0 and T, $V = 0$.

Evaluation of I_2

$$I_2 = \int_0^T \left[\int_{\mathbb{R}} \rho\mu \frac{\partial V}{\partial x} \,\mathrm{d}\, x \right] \mathrm{d}\, t = \int_0^T \left[\rho\mu V|_{\mathbb{R}} - \int_{\mathbb{R}} V \frac{\partial(\rho\mu)}{\partial x} \,\mathrm{d}\, x \right] \mathrm{d}\, t = -\int_{\mathbb{R}} \int_0^T \frac{\partial(\rho\mu)}{\partial x} V \,\mathrm{d}\, t \,\mathrm{d}\, x \tag{2.10}$$

where again we have used the fact that $\rho(X_t = \pm\infty, t | X_0, 0) = 0$.

Evaluation of I_3

$$\begin{aligned} I_3 &= \frac{1}{2} \int_0^T \left[\int_{\mathbb{R}} \rho\sigma^2 \frac{\partial^2 V}{\partial x^2} \,\mathrm{d}\, x \right] \mathrm{d}\, t = \frac{1}{2} \int_0^T \left[\rho\sigma^2 \frac{\partial V}{\partial x}\bigg|_{\mathbb{R}} - \int_{\mathbb{R}} \frac{\partial V}{\partial x} \frac{\partial(\rho\sigma^2)}{\partial x} \,\mathrm{d}\, x \right] \mathrm{d}\, t \\ &= -\frac{1}{2} \int_0^T \left[\int_{\mathbb{R}} \frac{\partial V}{\partial x} \frac{\partial(\rho\sigma^2)}{\partial x} \,\mathrm{d}\, x \right] \mathrm{d}\, t = -\frac{1}{2} \int_0^T \left[\frac{\partial(\rho\sigma^2)}{\partial x} V\bigg|_{\mathbb{R}} - \int_{\mathbb{R}} \frac{\partial^2(\rho\sigma^2)}{\partial x^2} V \,\mathrm{d}\, x \right] \mathrm{d}\, t \\ &= \frac{1}{2} \int_{\mathbb{R}} \int_0^T \frac{\partial^2(\rho\sigma^2)}{\partial x^2} V \,\mathrm{d}\, t \,\mathrm{d}\, x \end{aligned} \tag{2.11}$$

Since $\mathbb{E}[V(X_T, T) - V(X_0, 0)] \equiv 0$, we have

$$\int_{\mathbb{R}} \int_0^T \frac{\partial \rho}{\partial t} V \,\mathrm{d}\, t \,\mathrm{d}\, x + \int_{\mathbb{R}} \int_0^T \frac{\partial(\rho\mu)}{\partial x} V \,\mathrm{d}\, t \,\mathrm{d}\, x = \frac{1}{2} \int_{\mathbb{R}} \int_0^T \frac{\partial^2(\rho\sigma^2)}{\partial x^2} V \,\mathrm{d}\, t \,\mathrm{d}\, x \tag{2.12}$$

and

$$\frac{\partial \rho}{\partial t} = -\frac{\partial(\rho\mu)}{\partial x} + \frac{1}{2} \frac{\partial^2(\rho\sigma^2)}{\partial x^2} \tag{2.13}$$

Suppose the dynamics of X_t is instead specified by a geometric Brownian motion,

$$\mathrm{d}\, X_t \mu(X_t, t) X_t \,\mathrm{d}\, t + \sigma(X_t, t) X_t \,\mathrm{d}\, W_t \tag{2.14}$$

Then

$$\frac{\partial \rho}{\partial t} = -\frac{\partial(\rho\mu x)}{\partial x} + \frac{1}{2} \frac{\partial^2(\rho\sigma^2 x^2)}{\partial x^2} \tag{2.15}$$

This is the Fokker–Planck forward equation [80, 79].

Chapter 3

Statistical Finance

Chapter 4

Computational Finance

4.1 Linear Algebra

4.2 Numerical Analysis

4.3 Interpolation Methods

4.4 Root-finding Methods

4.4.1 The Bisection Method

4.4.2 The Newton–Raphson Method

4.4.3 The Secant Method

4.5 Optimization Methods

4.5.1 Linear Optimization

4.5.2 Non-linear Optimization

4.6 Software Engineering

Chapter 5

Financial Modeling

5.1 Overview of (Arbitrage) Pricing Strategies

5.2 The Black–Scholes Model and Its Variants

5.2.1 The Black–Scholes Model

5.2.2 The Back–Scholes Model with Constant Dividend Yield (BS-D)

5.2.3 The Black's Model

5.3 Forward Models

5.3.1 Funding Spread Models

5.3.2 Dividend Models

5.4 Short Rate Models

5.4.1 Overview

5.4.2 Ornstein–Uhlenbeck Process

5.4.3 Square-root Process

5.5 Volatility Models

5.5.1 Implied Volatility Surface

5.5.2 Local Volatility Model

$$C(K,T;S_0) = \int_K^\infty \phi(S_T,T;S_0)[S_T - K]\,\mathrm{d}\,S_T \tag{5.1}$$

Differentiating this twice with respect to K to obtain

$$\begin{aligned}\frac{\partial C^2}{\partial K^2} &= \frac{1}{\partial K^2}\left[\int_K^\infty \phi(S_T,T;S_0)S_T\,\mathrm{d}\,S_T - K\int_K^\infty \phi(S_T,T;S_0)\,\mathrm{d}\,S_T\right] \\ &= \frac{1}{\partial K}\left[-K\phi(K,T;S_0) - \int_K^\infty \phi(S_T,T;S_0)\,\mathrm{d}\,S_T + K\phi(K,T;S_0)\right] \\ &= -\frac{1}{\partial K}\left[\int_K^\infty \phi(S_T,T;S_0)\,\mathrm{d}\,S_T\right] \\ &= \phi(K,T;S_0) \end{aligned} \tag{5.2}$$

and differentiating this with respect to T we obtain

$$\frac{\partial C}{\partial T} = \int_K^\infty \left[\frac{\partial}{\partial T}\phi(S_T,T;S_0)\right](S_T-K)\,\mathrm{d}\,S_T \tag{5.3}$$

5.5.3 Stochastic Volatility Models

5.6 Jump Models

Chapter 6

Investment Management

6.1 Asset Classes

6.1.1 Fixed Income

6.1.2 Equities

6.1.3 Foreign Exchange

6.1.4 Credit Derivatives

6.1.5 Interest Rate Derivatives

Chapter 7

Risk Management

7.1 Risk Management Overview

7.1.1 Short-term Risk Management: Politics, Macroeconomics, Fundamental Analysis

7.1.2 Mid-term Risk Management: Technical Analysis, Sensitivity Analysis

7.1.3 Short-term Risk Management

7.2 Option Greeks

7.2.1 Delta

7.2.2 Gamma

7.2.3 Vega

7.2.4 Theta

7.2.5 Hedging

7.3 Correlation and Skew

7.3.1 Implied Volatility Surface

7.3.2 Correlation

7.3.3 Skew

7.4 Term Structure, Duration, and Convexity

7.4.1 Term Structure of Interest Rate

7.4.2 Duration

7.4.3 Convexity

Part II

Mathematics and Physics

Chapter 8

Calculus

8.1 Convergence Tests

There are five common techniques to test whether or not an infinite series is convergent. But first of all, a necessary condition:

Theorem 8.1.1 If the limit of the summand is undefined or nonzero, that is, $\lim_{n\to\infty} a_n \neq 0$, then the series $\sum_{n=1}^{\infty} a_n$ must diverge.

Theorem 8.1.2 Comparison Test. If $\{a_n\}, \{b_n\} > 0$, and the limit $\lim_{n\to\infty} \frac{a_n}{b_n}$ exists, is finite and is not zero, then $\sum_{n=1}^{\infty} a_n$ converges if and only if $\sum_{n=1}^{\infty} b_n$ converges.

Theorem 8.1.3 Integral Test. Let $f : [1, \infty) \to \mathbf{R}_+$ be a positive and monotone decreasing function such that $f(n) = a_n$. Then the series $\{a_n\}$ converges if and only if the integral $\int_1^{\infty} f(x)dx$ converges.

Theorem 8.1.4 Ratio Test. Suppose there exists r such that $\lim_{n\to\infty} |\frac{a_{n+1}}{a_n}| = r$. If $r < 1$, then the series converges. If $r > 1$, then the series diverges. If $r = 1$, the ratio test is inconclusive, and the series may converge or diverge.

Theorem 8.1.5 Root Test. Define $r = \limsup_{n\to\infty} \sqrt[n]{|a_n|}$. If $r < 1$, then the series converges. If $r > 1$, then the series diverges. If $r = 1$, the ratio test is inconclusive, and the series may converge or diverge.

Theorem 8.1.6 Alternating Series Test. If the alternating series $\sum_{n=1}^{\infty}(-1)^{n-1}b_n, (b_n > 0)$ satisfies

1. $b_{n+1} \leq b_n,$ for all n; and,
2. $\lim_{n\to\infty} b_n = 0$.

Then the series is convergent.

Theorem 8.1.7 A series is said to be absolutely convergent if $\sum_{i=1}^{\infty} |a_n|$ converges. Every absolutely convergent series is convergent. But not all convergent series are absolutely convergent. A convergent series that is not absolutely convergent is called conditionally convergent.

Chapter 9

Advanced Calculus

9.1 $\liminf$ and $\limsup$

Definition 9.1.1

$$\liminf_n A_n = \cup_{n=1}^{\infty} \cap_{i=n}^{\infty} A_i = \{x \mid x \in A_i \text{ eventually}\} \tag{9.1}$$

$$\limsup_n A_n = \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i = \{x \mid x \in A_i \text{ for infinitely many } i\} \tag{9.2}$$

The meaning of $\liminf$ can be seen by re-writing the above definition as: $x \in \liminf_n A_n$ if $\exists n \in \mathbb{N}$, s.t. $\forall i \geq n$ and $i \in \mathbb{N}$, $x \in A_i$. Hence the elements in $\liminf_n A_n$ are in all but (the first) finitely many sets, though the "first finitely many sets" may be different for different elements in $\liminf$. $\limsup$ can be best seen by examining its complement, according to the De Morgan's law.

Proposition 9.1.1

$$(\limsup_n A_n)^c = \liminf_n A_n^c \tag{9.3}$$

$$\liminf A_k \subset \limsup A_k \tag{9.4}$$

$$\limsup(A_k \cup B_k) = \limsup A_k \cup \limsup B_k \tag{9.5}$$

$$\liminf(A_k \cap B_k) = \liminf A_k \cap \liminf B_k \tag{9.6}$$

9.2 Lines in $\mathbb{R}^n$

Definition 9.2.1 Given a vector $\mathbf{p} \in \mathbb{R}^n$ and a nonzero vector $\mathbf{v} \in \mathbb{R}^n$, the set of all points $\mathbf{y} \in \mathbb{R}^n$ such that

$$\mathbf{y} = t\mathbf{v} + \mathbf{p}, \quad t \in \mathbb{R} \tag{9.7}$$

is called the *line* through $\mathbf{p}$ in the direction of $\mathbf{v}$.

Example 9.2.1 The shortest distance from a point $\mathbf{q} \in \mathbb{R}^n$ to a line L with equation $\mathbf{y} = t\mathbf{v} + \mathbf{p}$ is

$$\left\| (\mathbf{q} - \mathbf{p}) - \frac{(\mathbf{q} - \mathbf{p})^T \mathbf{v}}{\|\mathbf{v}\|^2} \mathbf{v} \right\| \tag{9.8}$$

9.3 Hyperplanes in $\mathbb{R}^n$

Definition 9.3.1 Suppose $\mathbf{n}$ is a normal vector for a hyperplan H through $\mathbf{p} \in \mathbb{R}^n$, then the normal equation for H is

$$\mathbf{n}^T(\mathbf{y} - \mathbf{p}) = 0 \tag{9.9}$$

If H is in $\mathbb{R}^3$, we can use cross-product $\times$ to obtain the normal vector given two vectors on the hyperplane.

Example 9.3.1 The shortest distance from a point $\mathbf{q} \in \mathbb{R}^n$ to a hyperplane H with equation $\mathbf{n}^T(\mathbf{y} - \mathbf{p}) = 0$ is

$$\left| \frac{\mathbf{n}^T(\mathbf{q} - \mathbf{p})}{\|\mathbf{n}\|} \right| \tag{9.10}$$

A hyperplane is a set satisfies $H = \{x : w^T x = b\}$. An equivalent form is $w^T(x - \frac{w}{\|w\|^2} b) = 0$, which suggests that the vector w is perpendicular to the hyperplane, called a normal vector.

Particularly, since $\frac{w^T w}{\|w\|^2} b = b$, we know $x_0 = \frac{w}{\|w\|} \frac{b}{\|w\|}$ is on the hyperplane. The x_0 is actually the projection of the origin, since w is orthogonal to the hyperplane and it is nothing but a scaled w on the hyperplane. Therefore, the shortest distance (along the direction of w) from the origin to the hyperplane is given by $\frac{b}{\|w\|}$ (could be negative, which means w is on the other side of the hyperplane).

In general, if a hyperplane is given by the equation $f(x) = w^T x - b = 0$, the distance from any arbitrary vector p to the hyperplane $w^T x = b$ is given by

$$\frac{f(p)}{\|w\|} = \frac{w^T p - b}{\|w\|}, \quad \text{if } p \text{ is on the opposite side of the origin} \tag{9.11}$$

$$-\frac{f(p)}{\|w\|} = -\frac{w^T p - b}{\|w\|}, \quad \text{if } p \text{ is on the same side of the origin} \tag{9.12}$$

Particularly, when $p = 0$ (the origin), the above becomes $\frac{b}{\|w\|}$, which agrees with our previous result.

Proof: Let's prove the first case. Suppose there is a vector x on the hyperplane, such that $p - x = d\frac{w}{\|w\|}$. Since w is orthogonal to the hyperplane, the scalar d is the distance we are after. Now, multiply both sides by w^T,

$w^T p - w^T x = d w^T \frac{w}{\|w\|}$
$w^T p - (w^T x - b) = d \frac{w^T w}{\|w\|} + b$
$w^T p = d\|w\| + b$
$d = \frac{w^T p - b}{\|w\|}$
The proof for the other case is similar.
■

9.4 The Real and Complex Number Systems

9.4.1 Definitions

Definition 9.4.1 If A is any set (whose elements may be numbers or any other objects), we write $x \in A$ to indicate that x is a memeber (or an element) of A. If x is not a memeber of A, we write: $x \notin A$.

Definition 9.4.2 Throughtout Chap. 1, the set of all rational numbers will be denoted by Q.

Definition 9.4.3 Suppose S is an ordered set, $E \subset S$, and E is bounded above. Suppose there exists an $\alpha \in S$ with the following properties:

(i) α is an upper bound of E.

(ii) If $\gamma < \alpha$ then γ is not an upper bound of E.

Then α is called the *least upper bound* of E [that there is at most one such α is clear from (ii)] or the *supremum* of E, and we write

$$\alpha = \sup E.$$

The *greatest lower bound*, or *infimum*, of a set E which is bounded below is defined in the same manner: The statement

$$\alpha = \inf E$$

means that α is a lower bound of E and that no β with $\beta > \alpha$ is a lower bound of E.

Definition 9.4.4 The extended real number system consists of the real field R and two symbols, $+\infty$ and $-\infty$. We preserve the original order in R, and define

$$-\infty < x < +\infty$$

for every $x \in R$.

9.5 Basic Topology

9.6 Finite, Countable, and Uncountable Sets

9.6.1 Definitions

Definition 9.6.1 Consider two sets A and B, whose elements may be any objects whatsoever, and suppose that with each element x of A there is associated, in some manner, an element of B, which we denote by $f(x)$. Then f is said to be a *function* from A to B (or a *mapping* of A into B). The set A is called the *domain* of f (we also say f is defined on A), and the elments $f(x)$ are called the *values* of f. The set of all values of f is called the *range* of f.

Definition 9.6.2 Let A and B be two sets and let f be a mapping of A into B. If $E \subset A$, $f(E)$ is defined to be the set of all elements $f(x)$, for $x \in E$. We call $f(E)$ the *image* of E under f. In this notation, $f(A)$ is the range of f. It is clear that $f(A) \subset B$. If $f(A) = B$, we say that f maps A *onto* B. (Note that, according to this usage, *onto* is more specific thatn *into*.)

Definition 9.6.3 If $E \subset B$ (E is not necessarily a subset of $f(A)$), $f^{-1}(E)$ denotes the set of all $x \in A$ such that $f(x) \in E$. We call $f^{-1}(E)$ the *inverse image* of E under f. If $y \in B$, $f^{-1}(y)$ is the set of all $x \in A$ such that $f(x) = y$. If, for each $y \in B$, $f^{-1}(y)$ consists of at most one element of A, then f is said to be a 1-1 (*one-to-one*) mapping of A into B. This may also be expressed as follows. f is a 1-1 mapping of A into B provided that $f(x_1) \neq f(x_2)$ whenever $x_1 \neq x_2, x_1 \in A, x_2 \in A$.

Definition 9.6.4 If there exists a 1-1 mapping of A *onto* B, we say that A and B can be put in *1-1 correspondence*, or that A and B have the same *cardinal number*, or, briefly, that A and B are *equivalent*, and we write $A \sim B$. This relation clearly has the following properties:

It is reflexive: $A \sim A$

It is symmetric: If $A \sim B$, then $B \sim A$

It is transitive: If $A \sim B$ and $B \sim C$, then $A \sim C$

Any relation with theses three properties is called an *equivalence relation*.

Definition 9.6.5 For any positive integer n, let J_n be the set whose elements are the integers $1, 2, \cdots, n$; let J be the set consisting of all positive integers. For any set A, we say:

(a) A is *finite* if $A \sim J_n$ for some n (the empty set is also considered to be finite).

(b) A is *infinite* if A is not finite.

(c) A is *countable* if $A \sim J$.

(d) A is *uncountable* if A is neither finite nor countable.

(e) A is *at most countable* if A is finite or countable.

Countable sets are sometimes called *enumerable* or *denumerable*.

Definition 9.6.6 By a *sequence*, we mean a function f defined on the set J of all positive integers. If $f(n) = x_n$, for $n \in J$, it is customary to denote the sequence f by the symbol $\{x_n\}$, or sometimes by $x_1, x_2, x_3, \cdots$. The values of f, that is, the elements x_n, are called the *terms* of the sequence. If A is a set and if $x_n \in A$ for all $n \in J$, then $\{x_n\}$ is said to be a *sequence in A*, or a *sequence of elements of A*.

Definition 9.6.7 Let A and Ω be sets, and suppose that with each element α of A there is associated a subset of Ω which we denote by E_α. The set whose elements are the sets E_α will be denoted by $\{E_\alpha\}$. Instead of speaking of sets of sets, we shall sometimes speak of a collection of sets, or a family of sets. The *union* of the sets E_α is defined to be the set S such that $x \in S$ if and only if $x \in E_\alpha$ for at least one $\alpha \in A$. We use the notation

$$S = \bigcup_{\alpha \in A} E_\alpha.$$

The *intersection* of the sets E_α is defined to be the set P such that $x \in P$ if and only if $x \in E_\alpha$ for every $\alpha \in A$. We use the notation

$$P = \bigcap_{\alpha \in A} E_\alpha.$$

9.6.2 Theorems

Theorem 9.6.1 A is infinite if and only if A is equivalent to one of its proper subsets.

Theorem 9.6.2 Every infinite subset of a countable set A is countable.

Theorem 9.6.3 Let $\{E_n\}, n = 1, 2, 3, \cdots$, be a sequence of countable sets, and put

$$S = \bigcup_{n=1}^{\infty} E_n.$$

Then S is countable.

Theorem 9.6.4 Let A be a countable set, and let B_n be the set of all n-tuples $(a_1, \cdots, a_n)$, where $a_k \in A (k = 1, \cdots, n)$, and the elements $a_1, \cdots, a_n$ need not be distinct. Then B_n is countable.

Corollary 9.6.1 The set of all rational numbers is countable.

Theorem 9.6.5 Let A be the set of all sequences whose elements are the digits 0 and 1. This set A is uncountable.

9.7 Metric Spaces

9.7.1 Definitions

Definition 9.7.1 A set X, whose elements we shall call *points*, is said to be a *metric space* if with any two points p and q of X there is associated a real number $d(p,q)$, called the *distance* from p to q, such that

(a) $d(p,q) > 0$ if $p \neq q$; $d(p,q) = 0$;

(b) $d(p,q) = d(q,p)$;

(c) $d(p,q) \leq d(p,r) + d(r,q)$, for any $r \in X$.

Any function with these three properties is called a *distance function*, or a *metric*.

Definition 9.7.2

(a) By the *segment* (a,b) we mean the set of all real numbers x such that $a < x < b$.

(b) By the *interval* $[a,b]$ we mean the set of all real numbers x such that $a \leq x \leq b$.

(c) Occasionally we shall also encounter "half-open intervals" $[a,b)$ and $(a,b]$; the first consist of all x such that $a \leq x < b$, the second of all x such that $a < x \leq b$.

(d) If $a_i < b_i$ for $i = 1, \cdots, k$, the set of all points $\mathbf{x} = (x_1, \cdots, x_k)$ in R^k whose coordinates satisfy the inequalities $a_i \leq x_i \leq b_i (1 \leq i \leq k)$ is called a *k-cell*.

(e) If $\mathbf{x} \in R^k$ and $r > 0$, the *open* (or *closed*) *ball* B with center at $\mathbf{x}$ and radius r is defined to be the set of all $y \in R^k$ such that $|\mathbf{y} - \mathbf{x}| < r$ (or $|\mathbf{y} - \mathbf{x}| < r$).

Definition 9.7.3 We call a set $E \subset R^k$ *convex* if

$$\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \in E$$

whenever $\mathbf{x} \in E, \mathbf{y} \in E$, and $0 < \lambda < 1$.

Definition 9.7.4 Let X be a metric space. All points and sets mentioned below are understood to be elements and subsets of X.

(a) A *neighborhood* of p is a set $N_r(p)$ consisting of all q such that $d(p,q) < r$, for some $r > 0$. The number r is called the *radius* of $N_r(p)$.

(b) A point p is a limit point of the set E if every neighborhood of p contains a point $q \neq p$ such that $q \in E$.

(c) E is *closed* if every limit point of E is a point of E.

(d) A point p is an *interior* point of E if there is a neighborhood N of p such that $N \subset E$.

(e) E is *open* if every point of E is an interior point of E.

(f) The *complement* of E (denoted by E^c) is the set of all points $p \in X$ such that $p \notin E$.

(g) E is *perfect* if E is closed and if every point of E is a limit point of E.

(h) E is *bounded* if there is a real number M and a point $q \in X$ such that $d(p, q) < M$ for all $p \in E$.

(i) E is *dense* in X if every point of X is a limit point of E, or a point of E (or both).

Definition 9.7.5 If X is a metric space, if $E \subset X$, and if E' denotes the set of all limit points of E in X, then the *closure* of E is the set $\bar{E} = E \cup E'$.

9.7.2 Theorems

Theorem 9.7.1

(a) Balls are convex.

(b) K-cells are convex.

Theorem 9.7.2 Every neighborhood is an open set.

Theorem 9.7.3 If p is a limit point of a set E, then every neighborhood of p contains infinitely many points of E.

Corollary 9.7.1 A finite point set has no limit points.

Theorem 9.7.4 Let $\{E_n\}$ be a (finite or infinite) collection of sets E_n. Then

$$\left(\bigcup_\alpha E_\alpha\right)^c = \bigcap_\alpha (E_\alpha^c)\,.$$

Theorem 9.7.5 A set F is closed if and only if its complement is open.

Theorem 9.7.6

(a) For any collection $\{G_n\}$ of open sets, $\bigcup_n G_n$ is open.

(b) For any collection $\{F_n\}$ of closed sets, $\bigcap_n F_n$ is closed.

(c) For any finite collection $G_1, \cdots, G_n$ of open sets, $\bigcap_{i=1}^n G_i$ is open.

(d) For any finite collection $F_1, \cdots, F_n$ of closed sets, $\bigcup_{i=1}^n F_i$ is closed.

Theorem 9.7.7 If X is a metric space and $E \subset X$, then

(a) $\bar{E}$ is closed,

(b) $E = \bar{E}$ if and only if E is closed,

(c) $\bar{E} \subset F$ for every closed set $F \subset X$ such that $E \subset F$.

By (a) and (c), $\bar{E}$ is the smallest closed subset of X that contains E,

Theorem 9.7.8 Let E be a nonempty set of real numbers which is bounded above. Let $y = supE$. Then $y \in \bar{E}$. Hence $y \in E$ if E is closed.

Theorem 9.7.9 Suppose $Y \subset X$. A subset E of Y is open relative to Y is and ony if $E = Y \cap G$ for some open subset G of X.

9.8 Compact Sets

9.8.1 Definitions

Definition 9.8.1 By an *open cover* of a set E in a metric space X we mean a collection $\{G_\alpha\}$ of open subsets of X such that $E \subset \bigcup_\alpha G_\alpha$.

9.8.2 Theorems

Theorem 9.8.1 Compact subsets of metric spaces are closed.

Theorem 9.8.2 Cloased subsets of compact sets are compact.

Theorem 9.8.3 If F is closed and K is compact, the n$F \cap K$ is compact.

Theorem 9.8.4 If $\{K_\alpha\}$ is a collection of compact subsets of a metric space X such that the intersection of every finite subcollection of $\{K_\alpha\}$ is nonempty, then $\cap K_\alpha$ is nonempty.

Theorem 9.8.5 If E is an infinite subset of a compact set K, then E has a limit point in K.

Theorem 9.8.6 If $\{I_n\}$ is a sequence of intervals in R^1, such that $I_n \supset I_{n+1}$ $(n = 1, 2, 3, \cdots)$, then $\cap_{n=1}^{\infty} I_n$ is not empty.

Theorem 9.8.7 Let k be a positive integer. If $\{I_n\}$ is a sequence of k-cells such that $I_n \supset I_{n+1}$ $(n = 1, 2, 3, \cdots)$, then $\cap_{n=1}^{i} nftyI_n$ is not empty.

Theorem 9.8.8 Every k-cell is compact.

Theorem 9.8.9 If a set E in R^k has one of the following three properties, then it has the other two:

(a) E is closed and bounded.

(b) E is compact.

(c) Every infinite subset of E has a limit point in E.

Theorem 9.8.10 Every bounded infinite subset of R^k has a limit point in R^k.

9.9 Perfect Sets

9.9.1 Theorems

Theorem 9.9.1 Let P be a nonempty perfect set in R^k. Then P is uncountable.

Corollary 9.9.1 Every interval $[a, b]$ $(a < b)$ is uncountable. In particular, the set of all real numbers is uncountable.

9.10 Connected Sets

9.10.1 Definitions

9.10.2 Theorems

Theorem 9.10.1 A subset E of the ral line R^1 is connected if and only if it has the following property: If $x \in E$, $y \in E$, and $x < z < y$, then $z \in E$.

9.11 Numerical Sequences and Series

9.12 Convergent Sequences

9.12.1 Definitions

Definition 9.12.1 The sequence $\{p_n\}$ is said to be *bounded* if its range is bounded.

9.12.2 Thorems

Theorem 9.12.1 Let $\{p_n\}$ be a sequence in a metric space X.

(a) $\{p_n\}$ converges to $p \in X$ if and only if every neighborhood of p contains p_n for all but finitely many n.

(b) If $p \in X, p' \in X$, and if $\{p_n\}$ converges to p and to p', then $p' = p$.

(c) If $\{p_n\}$ converges, then $\{p_n\}$ is bounded.

(d) If $E \subset X$ and if p is a limit point

Theorem 9.12.2 Suppose $\{s_n\}, \{t_n\}$ are complex sequences, and $\lim_{n\to\infty}\{s_n\} = s$ and $\lim\{t_n\} = t$. Then,

(a) $\lim_{n\to\infty}(s_n + t_n) = s + t$;

(b) $\lim_{n\to\infty}(cs_n) = cs, \lim_{n\to\infty}(c + s_n) = c + s$, for all number c;

(c) $\lim s_n t_n = st$;

(d) $\lim \frac{1}{s_n} = \frac{1}{s}$, provided $s_n \neq 0 (n = 1, 2, 3, \cdots)$, and $s \neq 0$.

Theorem 9.12.3

(a) Suppose $\mathbf{x}_n \in R^k (n = 1, 2, 3, \cdots)$ and

$$\mathbf{x}_n = (\alpha_{1,n}, \cdots, \alpha_{k,n}.$$

Then $\{\mathbf{x}_n\}$ converges to $\mathbf{x} = (\alpha_1, \cdots, \alpha_k)$ if and only if

$$\lim_{n \to \infty} \alpha_{j,n} = \alpha_j.$$

(b) Suppose $\{\mathbf{x}_n\}, \{\mathbf{y}_n\}$ are sequences in R^k, $\{\beta_n\}$ is a sequence of real numbers, and $\mathbf{x}_n \to \mathbf{x}, \mathbf{y}_n \to \mathbf{y}, \beta_n \to \beta$. Then

$$\lim_{n \to \infty} (\mathbf{x}_n + \mathbf{y}_n) = \mathbf{x} + \mathbf{y}, \lim_{n \to \infty} (\mathbf{x}_n \cdot \mathbf{y}_n) = \mathbf{x} \cdot \mathbf{y}, \lim_{n \to \infty} \beta_n \mathbf{x}_n = \beta \mathbf{x}.$$

9.13 Subsequences

9.13.1 Definitions

Definition 9.13.1 Given a sequence $\{p_n\}$, consider a sequence $\{n_k\}$ of positive integers, such that $n_1 < n_2 < n_3 < \cdots$. Then the sequence $\{p_{n_i}\}$ is called a *subsequence* of $\{p_n\}$. If $\{p_{n_i}\}$ converges, its limit is called a *subsequential limit* of $\{p_n\}$.

9.13.2 Theorems

Theorem 9.13.1 $\{p_n\}$ converges to p if and only if every subsequence of $\{p_n\}$ converges to p.

Theorem 9.13.2

(a) If $\{p_n\}$ is a sequence in a compact metric space X, then some subsequence of $\{p_n\}$ converges to a point of X.

(b) Every bounded sequence in R^k contains a convergent subsequence.

Theorem 9.13.3 The subsequential limits of a sequence $\{p_n\}$ in a metric space X form a closed subset of X.

9.14 Cauchy Sequences

9.14.1 Definitions

Definition 9.14.1 A sequence $\{p_n\}$ in a metric space X is said to be a *Cauchy sequence* if for every $\epsilon > 0$ there is an integer N such that $d(p_n, p_m) < \epsilon$ if $n \geq N$ and $m \geq N$.

Definition 9.14.2 Let E be a nonempty subset of a metric space X, and let S be the set of all real numbers of the form $d(p,q)$, with $p \in E$ and $q \in E$. The sup of S is called the *diameter* of E.
If $\{p_n\}$ is a sequence in X and if E_N consists of the points $p_N, p_{N+1}, p_{N+2}, \cdots$, it is clear from the two preceding deffs that $\{p_n\}$ is a *Cauchy sequence if and only if*

$$\lim_{N\to\infty} \text{diam } E_N = 0.$$

Definition 9.14.3 A metric space in which every Cauchy sequence converges is said to be *complete*.

Definition 9.14.4 A sequence $\{s_n\}$ of real numbers is said to be

(a) *monotonically* increasing if $s_n \leq s_{n+1} (n = 1, 2, 3, \cdots)$;

(b) *monotonically* decreasing if $s_n \geq s_{n+1} (n = 1, 2, 3, \cdots)$;

9.14.2 Theorems

Theorem 9.14.1

(a) If $\bar{E}$ is the closure of a set E in a metric space X, then

$$\text{diam } \bar{E} = \text{diam } E.$$

(b) If K_n is a sequence of compact sets in X such that $K_n \supset K_{n+1} (n = 1, 2, 3, \cdots)$ and if

$$\lim_{n\to\infty} \text{diam } K_n = 0,$$

then $\bigcap_1^\infty K_n$ consists of exactly one point.

Theorem 9.14.2

(a) In any metric space X, every convergent sequence is a Cauchy sequence.

(b) If X is a compact metric space and if $\{p_n\}$ is a Cauchy sequence in X, then $\{p_n\}$ converges to some point of X.

(c) in R^k, every Cauchy sequence converges.

The fact that a sequence converges in R^k if and only it is a Cauchy sequence is usually called the *Cauchy criterion* for convergence.
This thm says that *all compact metric spaces and all Euclidean spaces are complete*. It implies also that *every closed subset of E of a complete metric space X is complete*.

Theorem 9.14.3 Suppose $\{s_n\}$ is monotonic. Then $\{s_n\}$ converges if and only if it is bounded.

9.15 Upper and Lower Limits

9.15.1 Definitions

Definition 9.15.1 Let $\{s_n\}$ be a sequence of real numbers with the following property: For every real M there is an interger N such that $n \geq N$ implies $s_n \geq M$. We then write

$$s_n \to +\infty.$$

Similarly, if for every real M there is an integer N such that $n \geq N$ implies $s_n \leq M$, we write

$$s_n \to -\infty.$$

Definition 9.15.2 Let $\{s_n\}$ be a sequence of real numbers. Let E be the set of numbers x (in the extended real number system) such that $s_{n_k} \to x$ for some subsequence $\{s_{n_k}\}$. This set E contains all subsequential limits as defined in Definition 9.13.1, plus possibly the numbers $+\infty, -\infty$.
We now recall Definition 9.4.3 and 9.4.4 and put

$$s^* = \sup E,$$

$$s_* = \inf E.$$

The numbers s^*, s_* are called the *upper* and *lower limits* of $\{s_n\}$; we use the notation

$$\limsup_{n\to\infty} s_n = s^*, \quad \liminf_{n\to\infty} s_n = s_*$$

9.15.2 Theorems

Theorem 9.15.1 Let$\{s_n\}$ be a sequence of real numbers. Let E and s^* have the same meaning as in Definition 9.15.2. Then s^* has the following two properties:

(a) $s^* \in E$

(b) If $x > s^*$, there is an integer N such that $n \geq N$ implies $s_n < x$.

Moreover, s^* is the only number with the properties (a) and (b).
Of course, an analogous result is true for s_*.

Theorem 9.15.2 If $s_n \leq t_n$ for $n \geq N$, where N is fixed, then

$$\liminf_{n\to\infty} s_n \leq \liminf_{n\to\infty} t_n,$$

$$\limsup_{n\to\infty} s_n \leq \limsup_{n\to\infty} t_n,$$

9.16 Some Special Sequences

9.16.1 Theorems

Theorem 9.16.1

(a) If $p > 0$, then $\lim_{n\to\infty} \frac{1}{n^p} = 0$.

(b) If $p > 0$, then $\lim_{n\to\infty} \sqrt[n]{p} = 1$.

(c) $\lim_{n\to\infty} \sqrt[n]{n} = 1$.

(d) If $p > 0$ and α is real, then $\lim_{n\to\infty} \frac{n^\alpha}{(1+p)^n} = 0$.

(e) If $|x| < 1$, then $\lim_{n\to\infty} x^n = 0$.

9.17 Series

9.17.1 Definitions

Definition 9.17.1 Given a sequence $\{a_n\}$, we use the notation

$$\sum_{n=p}^{q} a_n \quad (p \le q)$$

to denote the sum $a_p + a_{p+1} + \cdots + a_q$. With $\{a_n\}$ we associate a sequence $\{s_n\}$, where

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

For $\{s_n\}$ we also use the symbolic expression

$$a_1 + a_2 + a_3 + \cdots$$

or, more concisely,

$$\sum_{n=1}^{\infty} a_n.$$

The above symbol we call an *infinite series*, or just a *series*. The numbers $\{s_n\}$ are called the *partial sums* of the series. If $\{s_n\}$ converges to s, we say that the series *converges*, and write

$$\sum_{n=1}^{\infty} a_n = s.$$

The number s is called the sum of the series; but it should be clearly understood that *s is the limit of a sequence of sums*, and is not obtained simply by addition. If $\{s_n\}$ diverges, the series is said to diverge.

9.17.2 Theorems

Theorem 9.17.1 $\sum a_n$ converges if and only if for every $\epsilon > 0$ there is an integer N such that

$$|\sum_{k=m}^{m} a_k| \leq \epsilon$$

if $m \geq n \geq N$.

Theorem 9.17.2 If $\sum a_n$ converges, then $\lim_{n\to\infty} a_n = 0$.

Theorem 9.17.3 A series of nonnegative terms converges if and only if its partial sums form a bounded sequence.

Theorem 9.17.4

(a) If $|a_n| \leq c_n$ for $n \geq N_0$, where N_0 is some fixed integer, and if $\sum c_n$ converges, then $\sum a_n$ converges.

(b) If $a_n \geq d_n \geq 0$ for $n \geq N_0$, and if $\sum d_n$ diverges, then $\sum a_n$ diverges.

9.18 Series of Nonnegative Terms

9.18.1 Theorems

Theorem 9.18.1 If $0 \leq x < 1$, then

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

If $x \geq 1$, the series diverges.

Theorem 9.18.2 Suppose $a_1 \geq a_2 \geq a_3 \geq \cdots \geq 0$. Then the series $\sum_{n=1}^{\infty} a_n$ converges if and only if the series

$$\sum_{k=0}^{\infty} 2^k a_{2^k} = a_1 + 2a_2 + 4a_4 + 8a_8 + \cdots$$

converges.

Theorem 9.18.3 $\sum \frac{1}{n^p}$ converges if $p > 1$ and diverges if $p \leq 1$.

Theorem 9.18.4 If $p > 1$,

$$\sum_{n=2}^{\infty} \frac{1}{n(\log n)^p}$$

converges; if $p \leq 1$, the series diverges.

9.19 The Number e

9.19.1 Definitions

Definition 9.19.1

$$e = \sum_{n=0}^{\infty} \frac{1}{n!}$$

9.19.2 Theorems

Theorem 9.19.1

$$\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e.$$

Theorem 9.19.2 e is irrational.

9.20 The Root and Ratio Tests

9.20.1 Theorems

Theorem 9.20.1 (Root Test) Given $\sum a_n$, put $\alpha = \limsup_{n\to\infty} \sqrt[n]{|a_n|}$. Then

(a) if $\alpha < 1$, $\sum a_n$ converges;

(b) if $\alpha > 1$, $\sum a_n$ diverges;

(c) if $\alpha = 1$, the test gives no information.

Theorem 9.20.2 (Ratio Test) The series $\sum a_n$

(a) converges if $\limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| < 1$,

(b) diverges if $\left|\frac{a_{n+1}}{a_n}\right| \geq 1$ for all $n \geq n_0$, where n_0 is some fixed integer.

Theorem 9.20.3 For any sequence $\{c_n\}$ of positive numbers,

$$\liminf_{n\to\infty} \frac{c_{n+1}}{c_n} \leq \liminf_{n\to\infty} \sqrt[n]{c_n},$$

$$\limsup_{n\to\infty} \sqrt[n]{c_n} \leq \limsup_{n\to\infty} \frac{c_{n+1}}{c_n}.$$

9.21 Power Series

9.21.1 Definitions

Definition 9.21.1 Given a sequence $\{c_n\}$ of complex numbers, the series

$$\sum_{n=0}^{\infty} c_n z^n$$

is called a *power series*. The numbers $\{c_n\}$ are called the *coefficients* of the series; z is a complex number.

9.21.2 Theorems

Theorem 9.21.1 Given the power series $\sum c_n z^n$, put

$$\alpha = \limsup_{n\to\infty} \sqrt[n]{|c_n|}, \quad R = \frac{1}{\alpha}.$$

(if $\alpha = 0, R = +\infty$; if $\alpha = +\infty, R = 0$.) Then $\sum c_n z^n$ converges if $|z| < R$, and diverges if $|z| > R$.

9.22 Summation by Parts

9.22.1 Theorems

Theorem 9.22.1 Given two sequences $\{a_n\}, \{b_n\}$, put

$$A_n = \sum_{k=0}^{n} a_k$$

if $n \geq 0$; put $A_{-1} = 0$. Then, if $0 \leq p \leq q$, we have

$$\sum_{n=p}^{q} a_n b_n = \sum_{n=p}^{q-1} A_n(b_n - b_{n+1}) + A_q b_q - A_{p-1} b_p.$$

Theorem 9.22.2 Suppose

(a) the partial sums A_n of $\sum a_n$ form a bounded sequences;

(b) $b_0 \geq b_1 \geq b_2 \geq \cdots$;

(c) $\lim_{n\to\infty} b_n = 0$.

Theorem 9.22.3 Suppose

(a) $|c_1| \geq |c_2| \geq |c_3| \geq \cdots$;

(b) $c_{2m-1} \geq 0, c_{2m} \leq 0 \quad (m = 1, 2, 3, \cdots)$;

(c) $\lim_{n\to\infty} c_n = 0$.

Then $\sum c_n$ converges.

Theorem 9.22.4 Suppose the radius of convergence of $\sum c_n z^n$ is 1, and suppose $c_0 \geq c_1 \geq c_2 \geq \cdots, \lim_{n\to\infty} c_n = 0$. Then $\sum c_n z^n$ converges at every point on the circle $|z| = 1$, except possibly at $z = 1$.

9.23 Absolute Convergence

9.23.1 Definitions

Definition 9.23.1 The series $\sum a_n$ is said to *converge absolutely* if the series $\sum |a_n|$ converges.

Definition 9.23.2 If $\sum a_n$ converges, but $\sum |a_n|$ diverges, we say that $\sum a_n$ converges *nonabsolutely*.

9.23.2 Theorems

Theorem 9.23.1 If $\sum a_n$ converges absolutely, then $\sum a_n$ converges.

9.24 Addition and Multiplication of Series

9.24.1 Definitions

Definition 9.24.1 Given $\sum a_n$ and $\sum b_n$, we put

$$c_n = \sum_{k=0}^{n} a_k b_{n-k} \quad (n = 0, 1, 2, \cdots)$$

and call $\sum c_n$ the *product* of the two given series.

9.24.2 Theorems

Theorem 9.24.1 If $\sum a_n = A$, and $\sum b_n = B$, then $\sum(a_n + b_n) = A + B$, and $\sum ca_n = cA$, for any fixed c.

Theorem 9.24.2 Suppose

(a) $\sum_{n=0}^{\infty} a_n$ converges absolutely,

(b) $\sum_{n=0}^{\infty} a_n = A$,

(c) $\sum_{n=0}^{\infty} b_n = B$,

(d) $c_n = \sum_{k=0}^{n} a_k b_{n-k} \quad (n = 0, 1, 2, \cdots)$.

Then

$$\sum_{n=0}^{\infty} c_n = AB.$$

That is, the product of two convergent series converges, and to the right value, if at least one of the two series converges absolutely.

Theorem 9.24.3 If the series $\sum a_n, \sum b_n, \sum c_n$ converge to A, B, C, and $c_n = a_0 b_n + \cdots + a_n b_0$ then $C = AB$.

9.25 Rearrangements

9.25.1 Definitions

Definition 9.25.1 Let $\{k_n\}, n = 1, 2, 3, \cdots$, be a sequence in which every positive integer appears once and only once (that is, $\{k_n\}$ is a 1-1 function from J onto J, in the notation of Definition 9.6.2). Putting

$$a'_n = a_{k_n} \quad (n = 1, 2, 3, \cdots),$$

we say that $\sum a'_n$ is a *rearrangement* of $\sum a_n$.

9.25.2 Theorems

Theorem 9.25.1 Let $\sum a_n$ be a series of real numbers which converges, but not absolutely. Suppose

$$-\infty \leq \alpha \leq \beta \leq \infty.$$

Then there exist a rearrangement $\sum a'_m$ with partial sums s'_n such that

$$\liminf_{n\to\infty} s'_n = \alpha, \quad \limsup_{n\to\infty} s'_n = \beta.$$

Theorem 9.25.2 If $\sum a_n$ is a series of complex numbers which converges absolutely, then every rearrangement of $\sum a_n$ converges, and they all converges to the same sum.

9.26 Continuity

9.27 Limits of Functions

9.27.1 Definitions

9.27.2 Theorems

Theorem 9.27.1 Let $X, Y, E, f,$ and p be as in Definition **??**. Then

$$\lim_{x\to p} f(x) = q$$

if and only if

$$\lim_{n\to\infty} f(p_n) = q$$

for every sequence $\{p_n\}$ in E such that

$$p_n \neq p, \quad \lim_{n\to\infty} p_n = p.$$

Corollary 9.27.1 If f has a limit at p, this limit is unique.

Theorem 9.27.2 Suppose $E \subset X$, a metric space, p is a limit point of E, f and g are complex functions on E, and

$$\lim_{x\to p} f(x) = A, \quad \lim_{x\to p} g(x) = B.$$

Then

(a) $\lim_{x\to p}(f+g)(x) = A + B$;

(b) $\lim_{x\to p}(fg)(x) = AB$;

(c) $\lim_{x\to p}(\frac{f}{g})(x) = \frac{A}{B}$, *if* $B \neq 0$.

9.28 Continuous Functions

9.28.1 Definitions

Definition 9.28.1 Suppose X and Y are metric spaces, $E \subset X$, $p \in E$, and f maps E into Y. Then f is said to be *continuous at p* if for every $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\delta_Y(f(x), f(p)) < \epsilon$$

for all points $x \in E$ for which $d_X(x, p) < \delta$.

9.28.2 Theorems

Theorem 9.28.1 Suppose X, Y, Z are metric spaces, $E \subset X$, f maps E into Y, g maps the range of f, $f(E)$, into Z, and h is the mapping of E into Z defined by

$$h(x) = g(f(x)) \quad (x \in E).$$

If f is continuous at a point $p \in E$ and if g is continuous at the point $f(p)$, then h is continuous at p.

Theorem 9.28.2 A mapping f of a metric space X into a metric space Y is onctinuous on X if and only if $f^{-1}(V)$ is open in X for every open set V in Y.

Corollary 9.28.1 A mapping f of a metric space X into a metric space Y is continuous if and only if $f^{-1}(C)$ is closed in X for every closed set C in Y.

Theorem 9.28.3 Let f and g be complex continuous functions on a metric space X. Then $f+g, fg,$ and f/g are continuous on X.

Theorem 9.28.4

(a) Let $f_1, \cdots, f_k$ be real functions on a metric space X, and let $\mathbf{f}$ be the mapping of X into R^k defined by

$$\mathbf{f}(x) = (f_1(x), \cdots, f_k(x)) \quad (x \in X);$$

then $\mathbf{f}$ is continuous if and only if each of the functions $f_1, \cdots, f_k$ is continuous.

(b) if $\mathbf{f}$ and $\mathbf{g}$ are continuous mappings of X into R^k, then $\mathbf{f}+\mathbf{g}$ and $\mathbf{f}\cdot\mathbf{g}$ are continuous on X.

9.29 Continuity and Compactness

9.29.1 Definitions

Definition 9.29.1 A mapping $\mathbf{f}$ of a set E into R^k is said to be *bounded* if there is a real number M such that $|\mathbf{f}(x)| \leq M$ for all $x \in E$.

Definition 9.29.2 Let f be a mapping of a metric space X into a metric space Y. We say that f is *uniformly continuous* on X if for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$d_Y(f(p), f(q)) < \epsilon$$

for all p and q in X for which $d_X(p,q) < \delta$.

9.29.2 Theorems

Theorem 9.29.1 Suppose f is a continuous mapping of a compact metric space X into a metric space Y. Then $f(X)$ is compact.

Theorem 9.29.2 If $\mathbf{f}$ is a continuous mapping of a compact metric space X into R^k, then $\mathbf{f}(X)$ is closed and bounded. Thus, $\mathbf{f}$ is bounded.

Theorem 9.29.3 Suppose f is a continuous 1-1 mapping of a compact metric space X onto a metric space Y. Then the inverse mapping f^{-1} defined on Y by

$$f^{-1}(f(x)) = x \quad (x \in X)$$

is a continuous mapping of Y onto X.

Theorem 9.29.4 Let f be a continuous mapping of a compact metric space X into a metric space Y. Then f is uniformly continuous on X.

Theorem 9.29.5 Let E be a noncompact set in R^1. Then

(a) there exists a continuous function on E which is not bounded;

(b) there exists a continuous and bounded function on E which has no maximum. If, in addition, E is bounded, then

(c) there exists a continuous function on E which is not uniformly continuous.

9.30 Continuity and Connectedness

9.30.1 Theorems

Theorem 9.30.1 If f is a continuous mapping of a metric space X into a metric space Y, and if E is a connected subset of X, then $f(E)$ is connected.

Theorem 9.30.2 Let f be a continuous real function on the interval $[a, b]$. If $f(a) < f(b)$ and if c is a number such that $f(a) < c < f(b)$, then there exists a point $x \in (a, b)$ such that $f(x) = c$.

9.31 Discontinuities

9.31.1 Definitions

Definition 9.31.1 Let f be defined on (a, b). Consider any point x such that $a \leq x < b$. We write

$$f(x+) = q$$

if $f(t_n) \to q$ as $n \to \infty$, for all sequences $\{t_n\}$ in (x, b) such that $t_n \to x$. To obtain the deff of $f(x-)$, for $a < x \leq b$, we restrict ourselves to sequences $\{t_n\}$ in (a, x). It is clear that any point x of (a, b), $\lim_{t \to x} f(t)$ exists if and only if

$$f(x+) = f(x-) = \lim_{t \to x} f(t).$$

9.32 Monotonic Functions

9.32.1 Definitions

Definition 9.32.1 Let f be real on (a, b). Then f is said to be *monotonically increasing* on (a, b) if $a < x < y < b$ implies $f(x) \leq f(y)$. If the last inequality is reversed, we obtain the deff of a *monotonically decreasing* function. The class of monotonic functions consists of both the increasing and the deceasing functions.

9.32.2 Theorems

Theorem 9.32.1 Let f be monotonically increasing on (a, b). Then $f(x+)$ and $f(x-)$ exist at every point of x of (a, b). More precisely,

$$\sup_{a<t<x} f(t) = f(x-) \leq f(x) \leq f(x+) = \inf_{x<t<b} f(t).$$

Furthermore, if $a < x < y < b$, then

$$f(x+) \leq f(y-).$$

Analogous results evidently hod for monotonically decreasing functions.

Theorem 9.32.2 Let f be monotonic on (a, b). Then the set of points of (a, b) at which f is discontinuous is at most countable.

9.33 Infinite Limits and Limits at Infinity

9.33.1 Definitions

Definition 9.33.1 For any real c, the set of real numbers x such that $x > c$ is called a neighborhood of $+\infty$ and is written $(c, +\infty)$. Similarly, the set $(-\infty, c)$ is a neighborhood of $-\infty$.

Definition 9.33.2 Let f be a real function defined on $E \subset R$. We say that

$$f(t) \to A \text{ as } t \to x,$$

where A and x are in the extended real number system, if for every neighborhood U of A there is a neighborhood V of x such that $V \cap E$ is not empty, and such that $f(t) \in U$ for all $t \in V \cap E, t \neq x$.

9.33.2 Theorems

Theorem 9.33.1 Let f and g be defined on $E \subset R$. Suppose

$$f(t) \to A, \quad g(t) \to B \text{ as } t \to x.$$

Then

(a) $f(t) \to A'$ implies $A' = A$.

(b) $(f + g)(t) \to A + B$,

(c) $(fg)(t) \to AB$,

(d) $(f/g)(t) \to A/B$,

provided the right member of (b), (c), and (d) are defined.

9.34 Differentiation

9.35 The Derivative of a Real Function

9.35.1 Definitions

Definition 9.35.1 Let f be defined (and real-valued) on $[a, b]$. For any $x \in [a, b]$ form the quotient

$$\phi(t) = \frac{f(t) - f(x)}{t - x} (a < t < b, t \neq x),$$

and define

$$f'(x) = \lim_{t \to x} \phi(t),$$

provided this limit exists in accordance with Definition **??**. We thus associate with the function f a function f' whose domain is the set of points x at which the limit exists; f' is called the *derivative* of f. If f' is defined at a point x, we say that f is *differentiable* at x. If f' is defined at every point of a set $E \subset [a, b]$, we say that f is differentiable on E.

9.35.2 Theorems

Theorem 9.35.1 Suppose f and g are defined on $[a, b]$ and are differentiable at a point $x \in [a, b]$. Then $f + g, fg,$ and f/g are differentiable at x, and

(a) $(f + g)'(x) = f'(x) + g'(x)$;

(b) $(fg)'(x) = f'(x)g(x) + f(x)g'(x)$;

(c) $(\frac{f}{g})'(x) = \frac{g(x)f'(x) - g'(x)f(x)}{g^2(x)}$

In (c), we assume of course that $g(x) \neq 0$.

Theorem 9.35.2 Suppose f is continuous on $[a, b]$, $f'(x)$ exists at some point $x \in [a, b]$, g is defined on an interval I which contains the range of f, and g is differentiable at the point $f(x)$. If

$$h(t) = g(f(t)) \quad (a \leq t \leq b),$$

then h is differentiable at x, and

$$h'(x) = g'(f(x))f'(x).$$

9.36 Mean Value Theorems

9.36.1 Definitions

Definition 9.36.1 Let f be a real function defined on a metric space X. We say that f has a *local maximum* at a point $p \in X$ if there exists $\delta > 0$ such that $f(q) \leq f(p)$ for all $q \in X$ with $d(p, q) < \delta$.

9.36.2 Theorems

Theorem 9.36.1 If f and g are continuous real functions on $[a, b]$ which are differentiable in (a, b), then there is a point $x \in (a, b)$ at which

$$[f(b) - f(a)]g'(x) = [g(b) - g(a)]f'(x).$$

Note that differentiability is not required at the endpoints.

Theorem 9.36.2 If f is a real continuous function on $[a, b]$ which is differentiable in (a, b), then there is a point $x \in (a, b)$ at which

$$f(b) - f(a) = (b - a)f'(x).$$

Theorem 9.36.3 Suppose f is differentiable in (a, b).

(a) If $f'(x) \geq 0$ for all $x \in (a, b)$, then f is monotonically increasing.

(b) If $f'(x) = 0$ for all $x \in (a, b)$, then f is constant.

(c) If $f'(x) \leq 0$ for all $x \in (a, b)$, then f is monotonically decreasing.

9.37 The Continuity of Derivatives

9.37.1 Theorems

Theorem 9.37.1 Suppose f is a real differentiable function on $[a, b]$ and suppose $f'(a) < \lambda < f'(b)$. Then there is a point $x \in (a, b)$ such that $f'(x) = \lambda$.

9.38 L'Hospital's Rule

9.38.1 Theorems

Theorem 9.38.1 Suppose f and g are real and differentiable in (a, b), and $g'(x) \neq 0$ for all $x \in (a, b)$, where $-\infty \leq a < b \leq +\infty$. Suppose

$$\frac{f'(x)}{g`(x)} \to A \; as \; x \to a.$$

If

$$f(x) \to 0 \; and \; g(x) \to 0 \; as \; x \to a,$$

or if

$$g(x) \to +\infty \; as \; x \to a,$$

then

$$\frac{f(x)}{g(x)} \to A \; as \; x \to a.$$

9.39 Derivatives of Higher Order

9.39.1 Definitions

Definition 9.39.1 If f has a derivative f' on an interval, and if f' is itself differentiable, we denote the derivative of f' b f'' and call f'' the second derivative of f. Continuing in this manner , we obtain functions

$$f, f', f'', f^{(3)}, \cdots, f^{(n)},$$

each of which is the derivative of the preceding one. $f^{(n)}$ is called the nth derivative, or the derivative of order n, of f.

9.40 Taylor's Theorem

9.40.1 Theorems

Theorem 9.40.1 Suppose f is a real function on $[a, b]$, n is a positive integer, $f^{(n-1)}$ is continuous on $[a, b]$, $f^{(n)}(t)$ exists for every $t \in (a, b)$. Let α, β be distincet points of $[a, b]$, and define

$$P(t) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\alpha)}{k!}(t-\alpha)^k.$$

Then there exists a point x between α and β such that

$$f(\beta) = P(\beta) + \frac{f^{(n)}(x)}{n!}(\beta-\alpha)^n.$$

9.41 Differentiation of Vector-valued Functions

9.41.1 Theorems

Theorem 9.41.1 Suppose $\mathbf{f}$ is a continuous mapping of $[a, b]$ into R^k and $\mathbf{f}$ is differentiable in (a, b). Then there exists $x \in (a, b)$ such that

$$|\mathbf{f}(b) - \mathbf{f}(a)| \leq (b-a)|\mathbf{f}'(x)|.$$

9.42 The Riemann-Stieltjes Integral

9.43 Definition and Existence of the Integral

9.43.1 Definitions

Definition 9.43.1 We say that the partition $P*$ is a *refinement* of P if $P* \supset P$

(that is, if every point of P is a point of $P*$). Given two partitions, P_1 and P_2, we say that $P*$ is their *common refinement* if $P* = P_1 \cup P_2$.

9.43.2 Theorems

Theorem 9.43.1 If P^* is a refinement of P, then

$$L(P, f, \alpha) \leq L(P^*, f, \alpha)$$

and

$$U(P^*, f, \alpha) \leq U(P, f, \alpha).$$

Theorem 9.43.2 $\underline{\int_a^b} f d\alpha \leq \overline{\int_a^b} f d\alpha$

Theorem 9.43.3 $f \in \mathbf{R}(\alpha)$ on $[a, b]$ if and only if for every $\epsilon > 0$ there exists a partition P such that

$$U(P, f, \alpha) - L(P, f, \alpha) < \epsilon.$$

Theorem 9.43.4

(a) If Theorem 9.43.3 holds for some P and some ϵ, then Theorem 9.43.3 holds (with the same ϵ) for every refinement of P.

(b) If Theorem 9.43.3 holds for $P = \{x_0, \cdots, x_n\}$ and if s_i, t_i are arbitrary points in $[x_{i-1}, x_i]$, then

$$\sum_{i=1}^{n} |f(s_i) - f(t_i)| \Delta\alpha_i < \epsilon.$$

(c) If $f \in \mathbf{R}(\alpha)$ and the hypotheses of (b) hold, then

$$|\sum_{i=1}^{n} |f(s_i) - f(t_i)| \Delta\alpha_i - \int_a^b f d\alpha| < \epsilon.$$

Theorem 9.43.5 If f is continuous on $[a, b]$ then $f \in \mathbf{R}(\alpha)$ on $[a, b]$.

Theorem 9.43.6 If f is monotonic on $[a, b]$, and if α is continuous on $[a, b]$, then $f \in \mathbf{R}(\alpha)$. (We still assume, of course, that α is monotonic.)

Theorem 9.43.7 Suppose f is bonded on $[a, b]$, f has only finitely many points of discontinuity on $[a, b]$, and α is continuous at every point at which f is discontinuous. Then $f \in \mathbf{R}(\alpha)$.

Theorem 9.43.8 Suppose $f \in \mathbf{R}(\alpha)$ on $[a, b]$, $m \leq f \leq M$, ϕ is continuous on $[m, M]$, and $h(x) = \phi(f(x))$ on $[a, b]$. Then $h \in \mathbf{R}(\alpha)$ on $[a, b]$.

9.44 Properties of the Integral

9.44.1 Definitions

Definition 9.44.1 The *unit step function* I is defined by

$$I(x) = \begin{cases} 0 & (x \leq 0) \\ 1 & (x > 0) \end{cases}$$

9.44.2 Theorems

Theorem 9.44.1 If $f \in \mathbf{R}(\alpha)$ and $g \in \mathbf{R}(\alpha)$ on $[a, b]$, then

(a) $fg \in \mathbf{R}(\alpha)$;

(b) $|f| \in \mathbf{R}(\alpha)$ and $\left|\int_a^b f d\alpha\right| \leq \int_a^b |f| d\alpha$.

Theorem 9.44.2 If $a < s < b$, f is bounded on $[a, b]$, f is continuous at s, and $\alpha(x) = I(x - s)$, then

$$\int_a^b f d\alpha = f(s).$$

Theorem 9.44.3 Suppose $c_n \geq 0$ for $1, 2, 3, \cdots$, $\sum c_n$ converges, $\{s_n\}$ is a sequence of distinct points in (a, b), and

$$\alpha(x) = \sum_{n=1}^{\infty} c_n I(x - s_n).$$

Let f be continuous on $[a, b]$. Then

$$\int_a^b f d\alpha = \sum_{n=1}^{\infty} c_n f(s_n).$$

Theorem 9.44.4 Assume α increases monotonically and $\alpha' \in \mathbf{R}$ on $[a, b]$. Let f be a bounded real function on $[a, b]$. Then $f \in \mathbf{R}(\alpha)$ if and only if $f\alpha' \in \mathbf{R}$. In that case,

$$\int_a^b f d\alpha = \int_a^b f(x)\alpha'(x) dx.$$

Theorem 9.44.5 Suppose ϕ is a strictly increasing continuous function that maps an interval $[A, B]$ onto $[a, b]$. Suppose α is monotonnically increasing on $[a, b]$ and $f \in \mathbf{R}(\alpha)$ on $[a, b]$. Define β and g on $[A, B]$ by

$$\beta(y) = \alpha(\phi(y)), \ g(y) = f(\phi(y)).$$

Then $g \in \mathbf{R}(\beta)$ and

$$\int_A^B g d\beta = \int_a^b f d\alpha.$$

9.45 Integration and Differentiation

9.45.1 Theorems

Theorem 9.45.1 Let $f \in \mathbf{R}$ on $[a, b]$. For $a \leq x \leq b$, put

$$F(x) = \int_a^x f(t)dt.$$

Then F is continuous on $[a, b]$; furthermore, if f is continuous at a point x_0 of $[a, b]$, then F is differentiable at x_0, and

$$F'(x_0) = f(x_0).$$

Theorem 9.45.2 If $f \in \mathbf{R}$ on $[a, b]$ and if there is a differentiable function F on $[a, b]$ such that $F' = f$, then

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

Theorem 9.45.3 Suppose F and G are differentiable functions on $[a, b]$, $F' = f \in \mathbf{R}$, and $G' = g \in \mathbf{R}$. Then

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

9.46 Sequences and Series of Functions

9.47 Discussion of Main Problem

9.47.1 Definitions

Definition 9.47.1 Suppose $\{f_n\}$, $n = 1, 2, 3, \cdots$, is a sequence of functions defined on a set E, and suppose that the sequence of numbers $\{f_n(x)$ converges for every $x \in E$. We can then define a function f by

$$f(x) = \lim_{n\to\infty} f_n(x) \quad (x \in E).$$

Under these circumstances we say that $\{f_n\}$ converges on E and that f is the *limit*, or the *limit function*, of $\{f_n\}$. Sometimes we shall use a more descriptive terminology and shall say that "$\{f_n\}$ converges to f *pointwise* on E" if the above holds. Similarly, if $\sum f_n(x)$ converges for every $x \in E$, and if we define

$$f(x) = \sum_{n=1}^{\infty} f_n(x) \quad (x \in E),$$

the function f is called the *sum* of the series $\sum f_n$.

9.48 Uniform Convergence

9.48.1 Definitions

Definition 9.48.1 We say that a sequence of functions $\{f_n\}, n = 1, 2, 3, \cdots$, converges *uniformly* on E to a function f if for every $\epsilon > 0$ there is an integer N such that $n \geq N$ implies

$$|f_n(x) - f(x)| \leq \epsilon$$

for all $x \in E$.

9.48.2 Theorems

Theorem 9.48.1 Suppose K is compact, and

(a) $\{f_n\}$ is a sequence of continuous functions on K,

(b) $\{f_n\}$ converges pointwise to a continuous function f on K,

(c) $f_n(x) \geq f_{n+1}(x)$ for all $x \in K, n = 1, 2, 3, \cdots$.

Then $f_n \to f$ uniformly on K.

Theorem 9.48.2 Supose

$$\lim_{n\to\infty} f_n(x) = f(x) \quad (x \in E).$$

Put

$$M_n = \sup_{x\in E} |f_n(x) - f(x)|.$$

Then $f_n \to f$ uniformly on E if and only if $M_n \to 0$ as $n \to \infty$.

Theorem 9.48.3 Suppose $\{f_n\}$ is a sequence of functions defined on E, and suppose

$$|f_n(x)| \leq M_n \quad (x \in E, n = 1, 2, 3, \cdots).$$

Then $\sum f_n$ converges uniformly on E if $\sum M_n$ converges.

9.49 Uniform Convergence and Continuity

9.49.1 Definitions

Definition 9.49.1 If X is a metric space, $\mathbf{C}(X)$ will denote the set of all complex-valued, continuous, bounded functions with domain X. We associate with each $f \in \mathbf{C}(X)$ its supreme norm

$$\|f\| = \sup_{x\in X} |f(x)|.$$

We also define the distance between $f \in \mathbf{C}(X)$ and $g \in \mathbf{C}(X)$ to be $\|f - g\|$.

9.49.2 Theorems

Theorem 9.49.1 Suppose $f_n \to f$ uniformly on a set E in a metric space. Let x be a limit point of E, and suppose that

$$\lim_{t \in x} f_n(t) = A_n \quad (n = 1, 2, 3, \cdots).$$

Then $\{A_n\}$ converges, and

$$\lim_{t \in x} f(t) = \lim_{n \to \infty} A_n.$$

In other words, the conclusion is that

$$\lim_{t \to x} \lim_{n \to \infty} f_n(t) = \lim_{n \to \infty} \lim_{t \to x} f_n(t).$$

Theorem 9.49.2 If $\{f_n\}$ is a sequence of continuous functions on E, and if $f_n \to f$ uniformly on E, then f is continuous on E.

Theorem 9.49.3 Suppose K is compact, and

(a) $\{f_n\}$ is a sequence of continuous functions on K,

(b) $\{f_n\}$ converges pointwise to a continuous function f on K,

(c) $f_n(x) \geq f_{n+1}(x)$ for all $x \in K, n = 1, 2, 3, \cdots$

Then $f_n \to f$ uniformly on K.

Theorem 9.49.4 The above metric makes $\mathbf{C}(X)$ into a complete metric space.

9.50 Uniform Convergence and Integration

9.50.1 Theorems

Theorem 9.50.1 Let α be monotonically increasing on $[a, b]$. Suppose $f_n \in \mathbf{R}(\alpha)$ on $[a, b]$, for $n = 1, 2, 3, \cdots$, and suppose $f_n \to f$ uniformly on $[a, b]$. Then $f \in \mathbf{R}(\alpha)$ on $[a, b]$, and

$$\int_a^b f d\alpha = \lim_{n \to \infty} \int_a^b f_n d\alpha.$$

(The existence of the limit is part of the conclusion.)

Corollary 9.50.1 If $f_n \in \mathbf{R}(\alpha)$ on $[a, b]$ and if

$$f(x) = \sum_{n=1}^{\infty} f_n(x) \quad (a \leq x \leq b),$$

the series converging uniformly on $[a, b]$, then

$$\int_a^b f d\alpha = \sum_{n=1}^{\infty} \int_a^b f_n d\alpha.$$

In other words, the series may be integrated term by term.

9.51 Uniform Convergence and Differentiation

9.51.1 Theorems

Theorem 9.51.1 Suppose $\{f_n\}$ is a sequence of functions, differentiable on $[a, b]$ and such that $\{f_n(x_0)\}$ converges for some point x_0 on $[a, b]$. If $\{f_n'\}$ converges uniformly on $[a, b]$, then $\{f_n\}$ converges uniformly on $[a, b]$, to a function f, and

$$f'(x) = \lim_{n\to\infty} f_n'(x) \quad (a \le x \le b).$$

Theorem 9.51.2 There exists a real continuous function on the real line which is nowhere differentiable.

9.52 Equicontinuous Families of Functions

9.52.1 Definitions

Definition 9.52.1 Let $\{f_n\}$ be a sequence of functions defined on a set E. We say that $\{f_n\}$ is *pointwise bounded* on E if the sequence $\{f_n(x)\}$ is bounded for every $x \in E$, that is, if there exists a finite-valued function ϕ defined on E such that

$$|f_n(x)| < \phi(x) \quad (x \in E, n = 1, 2, 3, \cdots).$$

We say that $\{f_n\}$ is *uniformly bounded* on E if there exists a number M such that

$$|f_n(x)| < M \quad (x \in E, n = 1, 2, 3, \cdots).$$

Definition 9.52.2 A family $\mathbf{F}$ of complex functions f defined on a set E in a metric space X is said to be *equicontinuous* on E if for every $\epsilon > 0$ there exists a $\delta > 0$ such that

$$|f(x) - f(y)| < \epsilon$$

whenever $d(x, y) < \delta, x \in E, y \in E,$ *and* $f \in \mathbf{F}$.

9.52.2 Theorems

Theorem 9.52.1 If $\{f_n\}$ is a pointwise bounded sequence of complex functions on a countable set E, then $\{f_n\}$ has a subsequence $\{f_{n_k}\}$ such that $\{f_{n_k}\}$ converges for every x inE.

Theorem 9.52.2 If K is a compact metric space, if $f_n \in \mathbf{C}(K)$ for $n = 1, 2, 3, \cdots$, and if $\{f_n\}$ converges uniformly on K, then $\{f_n\}$ is equicontinuous on K.

Theorem 9.52.3 If K is compact, if $f_n \in \mathbf{C}(K)$ for $n = 1, 2, 3, \cdots$, and if $\{f_n\}$ is pointwise bounded and equicontinuous on K, then

(a) $\{f_n\}$ is uniformly bounded on K,

(b) $\{f_n\}$ contains a uniformly convergent subsequence.

9.53 The Stone-Weierstrass Theorem

9.53.1 Theorems

Theorem 9.53.1 If f is a continuous complex function on $[a, b]$, there exists a sequence of polynomials P_n such that

$$\lim_{n\to\infty} P_n(x) = f(x)$$

uniformly on $[a, b]$. If f is real, the P_n may be taken real.

Corollary 9.53.1 For every interval $[-a, a]$ there is a sequence of real polynomials P_n such that $P_n(0) = 0$ and such that

$$\lim_{n\to\infty} P_n(x) = |x|$$

uniformly on $[-a, a]$.

9.54 Some Special Functions

9.55 Functions of Several Variables

9.56 The Contraction Principle

9.56.1 Definitions

9.56.2 Theorems

Theorem 9.56.1 If X is a complete metric space, and if ϕ is a contraction of X into X, then there exists one and only one $x \in X$ such that $\phi(x) = x$.

9.57 Exercises

9.57.1 Concept Questions

Problem 9.57.1 A sequence $\{a_n\}$ converges if and only if it is bounded.
- FALSE. $\{sin(n)\}$ is bounded but not convergent. However, if a sequence converges, then it is bounded. See Theorem **??**

Chapter 10

Real Analysis

10.1 Introduction

The core material of real analysis is that of Lebesgue integral, which extends the application of Riemann integral to a larger family of functions. The prerequisite of Lebesgue integral is measure theory. We begin from important concepts of sets, point topology, and the real number system, then continue with measurable functions before discussing Lebesgue integral.

10.2 Set Theory

10.3 Point Topology

10.4 Real Number System

The real number system can be characterized by three axioms: 1) the filed axiom, 2) the order axiom, and 3) the completeness axiom.

Of particular interest is the completeness axiom. Depending on the construction of real numbers, it can take the form of axioms (the completeness axiom), or a theorem from the construction. These include:

1. Lease upper bound property
2. Dedekind completeness
3. Cauchy completeness
4. Nested intervals theorem
5. Monotone convergence theorem
6. Bolzano-Weierstrass theorem

10.5 Measure Theory

10.6 Measurable Sets and Measurable Functions

10.7 Lebesgue Integration

10.8 Preface

Some good books to consider:

1. Linear Algebra Its Applications, by Strang
2. Linear Algebra Right, by Axler
3. Linear Algebra, by Lang
4. Finite Dimensional Spaces Mathematics Studies, by Halmos
5. Linear Algebra Problem Book, by Halmos
6. Linear Algebra and Its Applications, by Lax
7. http://joshua.smcvt.edu/linearalgebra/book.pdf
8. http://www.math.brown.edu/ treil/papers/LADW/book.pdf

10.9 Preliminaries

Definition 10.9.1 A **field** is a non-empty set F *closed* under two operations, usually called *addition* and *multiplication*[1], and denoted by $+$ and $\cdot$ respectively, such that the following *nine* axioms hold

(1-2). Associativity of addition and multiplication.

(3-4). Commutativity of addition and multiplication.

(5-6). Existence and uniqueness of additive and multiplicative identity elements.

(7-8). Existence and uniqueness of additive inverses and multiplicative inverses.

(9). Distributivity of multiplication over addition.

Definition 10.9.2 The characteristic of a ring R, $char(R)$, is the smallest positive integer n such that

$$\underbrace{1 + \cdots + 1}_{n \text{ summands}} = 0$$

Theorem 10.9.1 Any finite ring has nonzero characteristic.

[1]Subtraction and division are defined implicitly in terms of the inverse operations of addition and multiplication.

Chapter 11

Linear Algebra

11.1 Vector Space

Definition 11.1.1 A **vector space** over a field $\mathcal{F}$ is a *nonempty* set V together with the operations of addition $V \times V \to V$ and scalar multiplication $\mathcal{F} \times V \to V$ satisfying the following *eight* properties:

(-) Additive axioms. For every $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$, we have

(1) $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$

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

(3) $\mathbf{0} + \mathbf{u} = \mathbf{u} + \mathbf{0} = \mathbf{u}$, where $\mathbf{0} \in V$ is unique for all $\mathbf{u} \in V$

(4) $(-\mathbf{u}) + \mathbf{u} = \mathbf{u} + (-\mathbf{u}) = \mathbf{0}$, where $-\mathbf{u} \in V$ is unique for every $\mathbf{u} \in V$

(-) Multiplicative axioms. For every $\mathbf{u} \in V$ and scalars $a, b \in \mathcal{F}$, we have

(1) $1\mathbf{x} = \mathbf{x}$

(2) $(ab)\mathbf{x} = a(b\mathbf{x})$

(-) Distributive axioms. For every $\mathbf{u}, \mathbf{v} \in V$ and scalars $a, b \in \mathcal{F}$, we have

(1) a($\mathbf{u}$+$\mathbf{v}$) = a$\mathbf{u}$+ a$\mathbf{v}$

(2) (a+b)$\mathbf{u}$= a$\mathbf{u}$+ b$\mathbf{u}$

11.2 Subspaces

Definition 11.2.1 A subspace of $\mathcal{R}^n$ is any collection S of vectors in $\mathcal{R}^n$ such that

(1) The zero vector $\mathbf{0}$ is in S.

(2) If $\mathbf{u}$ and $\mathbf{v}$ are in S, then $\mathbf{u} + \mathbf{v}$ is in S. [1]

[1] S is closed under addition.

(3) If $\mathbf{u}$ is in S and c is a scalar, then $c\mathbf{u}$ is in S. [2]

Definition 11.2.2 Let S, T be two subspaces of $\mathcal{R}^n$. We say S is orthogonal to T if *every* vector in S is orthogonal to *every* vector in T. The subspace $\{\mathbf{0}\}$ is orthogonal to all subspaces. [3]

Definition 11.2.3 Let A be an $m \times n$ matrix.

(1) The *row space* of A is the subspace $row(A)$ of $\mathcal{R}^n$ spanned by the rows of A.

(2) The *column space* (or *range*) of A is the subspace $col(A)$ of $\mathcal{R}^m$ spanned by the columns of A.

11.2.1 Four Important Subspaces: the row, column, null, and left null space

Definition 11.2.4 Let A be an $m \times n$ matrix. The *null space* (or *kernel*) of A is the subspace of $\mathcal{R}^n$ consisting of solutions of the homogeneous linear system $A\mathbf{x} = \mathbf{0}$. It is denoted by *null(A)*.

Definition 11.2.5 A *basis* for a subspace S of $\mathcal{R}^n$ is a set of vectors in S that

(1) spans S and

(2) is linearly independent. [4]

Definition 11.2.6 If S is a subspace of $\mathcal{R}^n$, then the number of vectors in a basis for S is called the *dimension* of S, denoted *dim* S. [5]

Definition 11.2.7 The *rank* of a matrix A is the dimension of its row and column spaces and is denoted by *rank(A)*. [6]

Definition 11.2.8 The *nullity* of a matrix A is the dimension of its null space and is denoted by *nullity(A)*.

Theorem 11.2.1 The Rank Theorem. If A is an $m \times n$ matrix, then

$$rank(A) + nullity(A) = n$$

.

Theorem 11.2.2 If A is invertible, then A is a product of elementary matrices.

---

[2] S is closed under scalar multiplication.

[3] A line can be orthogonal to another line, or it can be orthogonal to a plane, but a plane cannot be orthogonal to a plane.

[4] It does not mean that they are orthogonal.

[5] The zero vector $\mathbf{0}$ is always a subspace of $\mathcal{R}^n$. Yet any set containing the zero vector is linearly dependent, so $\mathbf{0}$ cannot have a basis. We define *dim* $\mathbf{0}$ to be 0.

[6] The row and column spaces of a matrix A have the same dimension.

Theorem 11.2.3 Let A be an $m \times n$ matrix. Then $rank(A^T A) = rank(A)$.

Definition 11.2.9 Let S be a subspace of $\mathcal{R}^n$ and let $B = \{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ be a basis for S. Let $\mathbf{v}$ be a vector in S, and write $\mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k$. Then $c_1, \cdots, c_k$ are called the coordinates of $\mathbf{v}$ with respect to B, and the column vector

$$[\mathbf{v}]_B = [c_1, \cdots, c_k]^T$$

is called the coordinate vector of $\mathbf{v}$ with respect to B. [7]

Definition 11.2.10 A transformation $T : \mathcal{R}^n \to \mathcal{R}^m$ is called a linear transformation if

$$T(c_1\mathbf{v}_1 + c_2\mathbf{v}_2) = c_1T(\mathbf{v}_1) + c_2T(\mathbf{v}_2)$$

for all $\mathbf{v}_1, \mathbf{v}_2$ in $\mathcal{R}^n$ and scalars c_1, c_2.

11.3 Bases and Dimension

11.4 Coordinates

11.5 Linear Forms: One Vector as Argument

11.6 Bilinear and Quadratic Forms: Two Vectors as Argument

11.7 Jordan Canoical Forms

11.8 Eigenvalues and Eigenvectors

11.9 Definitions

Remark 11.9.1 eigenvectors are non-zero.

Definition 11.9.1 The set of all eigenvectors corresponding to the same eigenvalue, together with the zero vector, is called an *eigenspace*.

Definition 11.9.2 The characteristic polynomial of a matrix $\mathbf{A}$ of order n is

$$|\mathbf{A} - \lambda\mathbf{I}| = \prod_{i=1}^{n}(\lambda - \lambda_i) \tag{11.1}$$

Theorem 11.9.1 Every square matrix of order n has n eigenvalues, possibly complex and not necessarily all unique.

[7] This coordinate vector is unique.

Definition 11.9.3 The algebraic multiplicity $\mu_A(\lambda_i)$ of a eigenvalue λ_i is the multiplicity as a root of the characteristic polynomial.

Definition 11.9.4 The eigenspace E_{λ_i} associated with λ_i is defined as

$$E_{\lambda_i} = \{\mathbf{v} : (\mathbf{A} - \lambda_i \mathbf{I})\mathbf{v} = 0\} \tag{11.2}$$

Definition 11.9.5 The dimension of the eigenspace $\mathbf{E}_{\lambda_i}$ is referred to as the geometric multiplicity $\gamma_A(\lambda_i)$ of λ_i.

11.10 Vector Calculus

11.11 Inner Product (Dot Product)

Definition 11.11.1

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^T$$

Remark 11.11.1 The inner product is the trace of the outer product.

11.12 Outer Product

11.13 Cross Product

Definition 11.13.1

$$\mathbf{a} \times \mathbf{b} = \|\mathbf{a}\|\|\mathbf{b}\| sin(\theta)\mathbf{n}$$

It is also called the vector product.

11.14 Scalar Triple Product

11.15 Vector Triple Product

11.16 Line, Surface, and Volume Integrals

11.17 Integration of Vectors and Matrices

11.18 Matrix Calculus

11.19 Matrix Determinant

11.20 Kronecker Product and Vec

11.21 Hadamard Product and Diag

11.22 Matrix Exponential

11.23 Vector and Matrix Derivatives

Suppose $\mathbf{Y}_{m\times n}$ and $\mathbf{X}_{p\times q}$ are both matrices (scalars, vectors are of course special cases). The derivative of $\mathbf{Y}$ with respect to $\mathbf{X}$ involves $mnpq$ partial derivatives, $\left[\frac{\partial Y_{ij}}{\partial X_{kl}}\right]$, for $i = 1, \cdots, m; j = 1, \cdots, n; k = 1, \cdots, p; l = 1, \cdots, q$. This immediately poses a question: What is a convenient (or logic) way of arraying these partial derivatives - as a row vector, as a column vector, or as a matrix (which is a natural choice), and if the latter of what shape/order?

Two competing notational conventions can be distinguished by whether the index of the derivative (matrix) is majored by the numerator or the denominator.

1. Numerator layout, i.e. according to $\mathbf{Y}$ and $\mathbf{X}^T$. This is sometimes known as the Jacobian layout.

2. Denominator layout, i.e. according to $\mathbf{Y}^T$ and $\mathbf{X}$. This is sometimes known as the gradient layout. It is named so because the gradient under this layout is a usual column vector.

The transpose of one layout is the same as the other. We use the **numerator-layout** notation throughout the paper.

	Scalar	Vector	Matrix
Scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x} = [\frac{\partial y_i}{\partial x}]$	$\frac{d\mathbf{Y}}{\partial x} = [\frac{\partial y_{ij}}{\partial x}]$
Vector	$\frac{\partial y}{\partial \mathbf{x}} = [\frac{\partial y}{\partial x_j}]$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = [\frac{\partial y_i}{\partial x_j}]$	
Matrix	$\frac{\partial y}{d\mathbf{X}} = [\frac{\partial y}{\partial x_{ji}}]$		

The partials with respect to the numerator are laid out according to the shape Y while the partials with respect to the denominator are laid out according to the transpose of X. For example, $\partial y/\partial \mathbf{x}$ is a row vector[8] while $\partial \mathbf{y}/\partial x$ is a column vector.

Note:

1. derivative is a row vector; gradient is its transpose.
2. Hessian is the derivative of gradient.

11.24 Differentials

Example 11.24.1

$$d\mathbf{A} = \mathbf{A} - \mathbf{A} = \mathbf{0} \tag{11.3}$$

Example 11.24.2

$$d(\alpha \mathbf{X}) = \alpha(X + d\mathbf{X}) - \alpha \mathbf{X} = \alpha d\mathbf{X} \tag{11.4}$$

Example 11.24.3

$$d(\mathbf{X} + \mathbf{Y}) = [(\mathbf{X} + \mathbf{Y}) + d(\mathbf{X} + \mathbf{Y})] - (\mathbf{X} + \mathbf{Y}) = d\mathbf{X} + d\mathbf{Y} \tag{11.5}$$

Example 11.24.4

$$d(\mathrm{tr}(\mathbf{X})) = \mathrm{tr}(\mathbf{X} + d\mathbf{X}) - \mathrm{tr}(\mathbf{X}) = \mathrm{tr}(\mathbf{X} + d\mathbf{X} - \mathbf{X}) = \mathrm{tr}(d\mathbf{X}) \tag{11.6}$$

Example 11.24.5

$$d(\mathbf{XY}) = (\mathbf{X} + d\mathbf{X})(\mathbf{Y} + d\mathbf{Y}) - \mathbf{XY} = [\mathbf{XY} + \mathbf{X}d\mathbf{Y} + (d\mathbf{X})\mathbf{Y} + d\mathbf{X}d\mathbf{Y}] - \mathbf{XY} = \mathbf{X}d\mathbf{Y} + (d\mathbf{X})\mathbf{Y} \tag{11.7}$$

[8] We distinguish $\partial y/\partial \mathbf{x}$ and the gradient $\nabla_{\mathbf{x}} y$, which is the transpose of the former and hence a column vector.

Example 11.24.6

$$\mathbf{0} = \mathbf{dI} = d(\mathbf{X}\mathbf{X}^{-1}) = (d\mathbf{X})\mathbf{X}^{-1} + \mathbf{X}d\mathbf{X}^{-1} \tag{11.8}$$

$$d\mathbf{X}^{-1} = -\mathbf{X}^{-1}(d\mathbf{X})\mathbf{X}^{-1} \tag{11.9}$$

Another proof is:

$$\frac{\mathbf{A}^{-1}(x+h) - \mathbf{A}^{-1}(x)}{h} = \frac{\mathbf{A}^{-1}(x+h)[\mathbf{A}(x+h) - \mathbf{A}(x)]\mathbf{A}^{-1}(x)}{h}$$

Next, let's prove something not so trivial.

Proposition 11.24.1

$$d|\mathbf{X}| = |\mathbf{X}|\text{tr}(\mathbf{X}^{-1}d\mathbf{X}) \tag{11.10}$$

Proof First, we see that

$$\text{tr}(\mathbf{A}^T\mathbf{B}) = \sum_{i=1}^{n}\left(\sum_{j=1}^{n}(\mathbf{A}^T)_{ij}\mathbf{B}_{ji}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{A}_{ji}\mathbf{B}_{ji} = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{A}_{ij}\mathbf{B}_{ij} = \text{vec}(\mathbf{A})^T\text{vec}(\mathbf{B}) \tag{11.11}$$

which can be computed by first multiply $\mathbf{A}$ and $\mathbf{B}$ element-wise, and then sum all the elements in the resulting matrix (known as the *Frobenius inner product*)[9].

Next, applying the Laplace's formula

$$|\mathbf{X}| = \sum_j x_{ij} \cdot adj^T(\mathbf{X})_{ij} \tag{11.12}$$

we have,

$$d(|\mathbf{X}|) = \sum_i\sum_j \frac{\partial|\mathbf{X}|}{\partial x_{ij}}dx_{ij} \tag{11.13}$$

$$= \sum_i\sum_j \frac{\partial\{\sum_k x_{ik} \cdot adj^T(\mathbf{X})_{ik}\}}{\partial x_{ij}}dx_{ij} \qquad \text{(expand by row } i\text{)} \tag{11.14}$$

$$= \sum_i\sum_j \left\{\sum_k \frac{\partial x_{ik}}{\partial x_{ij}} \cdot adj^T(\mathbf{X})_{ik} + \sum_k x_{ik}\frac{\partial adj^T(\mathbf{X})_{ik}}{\partial x_{ij}}\right\}dx_{ij} \tag{11.15}$$

$$= \sum_i\sum_j adj^T(\mathbf{X})_{ij}dx_{ij} \qquad \left(\frac{\partial adj^T(\mathbf{X})_{ik}}{\partial x_{ij}} = 0, \forall k \neq j\right) \tag{11.16}$$

[9]The trace operator is a scalar function (of a matrix), that essentially turns matrices into vectors and computes a dot product between them.

(11.17)

Now, use $\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{A}_{ij}\mathbf{B}_{ij} = \mathrm{tr}(\mathbf{A}^T\mathbf{B})$, we have

$$d(|\mathbf{X}|) = tr(adj(X)d\mathbf{X}) \tag{11.18}$$

Since $\mathbf{X}$ is invertible, and $adj(\mathbf{X}) = |\mathbf{X}|\mathbf{X}^{-1}$, finally,

$$d(|\mathbf{X}|) = |\mathbf{X}|tr(\mathbf{X}^{-1}d\mathbf{X}) \tag{11.19}$$

11.25 Vector-by-vector Derivatives

The first two important identities are

$$\frac{\partial A\mathbf{x}}{\partial \mathbf{x}} = A \tag{11.20}$$

$$\frac{\partial \mathbf{x}^T A}{\partial \mathbf{x}} = A^T \tag{11.21}$$

In the numerator-layout, the major index of the resulting matrix is based on the numerator, so when A is on the left hand side of $\mathbf{x}$, the derivative is the same size as A, on the other hand, if A is on the right hand side of $\mathbf{x}$, it needs to be transposed.

Example 11.25.1 Suppose $a = a(\mathbf{x})$ is a scalar function and $\mathbf{u} = \mathbf{u}(\mathbf{x})$ a vector function.

$$\frac{\partial a\mathbf{u}}{\partial \mathbf{x}} = \frac{\partial a\mathbf{u}}{\partial a}\frac{\partial a}{\partial \mathbf{x}} + \frac{\partial a\mathbf{u}}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}} = \mathbf{u}\frac{\partial a}{\partial \mathbf{x}} + a\frac{\partial \mathbf{u}}{\partial \mathbf{x}} \tag{11.22}$$

Recall that $\frac{\partial a\mathbf{u}}{\partial a}$ is a row vector, and the chain rule is expanded from right to left, just as the composition of functions.

11.26 Derivatives of Vectors and Matrices

11.26.1 Derivatives of a Vector or Matrix with Respect to a Scalar

Let $\mathbf{A}$ be a matrix, as a matrix-valued function

$$\mathbf{A}(x) : \mathcal{R} \to \mathcal{R}^{m\times n} \tag{11.23}$$

For vector- and matrix-valued functions there is a further manifestation of the linearity of the derivative: Suppose that f is a fixed linear function defined on $\mathcal{R}^n$ and that $\mathbf{A}$ is a differentiable vector- or matrix-valued function. Then

$$f(\mathbf{A})' = f(\mathbf{A}') \tag{11.24}$$

A useful example is the trace of $\mathbf{A}$, which is the sum of the diagonal elements of $\mathbf{A}$ (differentiable real-valued functions)

$$tr(\mathbf{A})' = tr(\mathbf{A}') \tag{11.25}$$

Another example is the inner product of two vectors, where we have [10]

$$(\mathbf{a}^T\mathbf{b})' = \mathbf{a}'^T\mathbf{b} + \mathbf{a}^T\mathbf{b}' \tag{11.26}$$

11.27 Vector and Matrix Integrals

11.28 Some Intuitive Explanations

11.29 Eigenvalues and Singular Values

11.30 SVD, PCA, and Change of Basis

11.31 Special Square Matrices

11.32 Elementary Matrices

There are three types of elementary matrices: Row Switching, Row Multiplication, and Row Addition.

Remark 11.32.1 Left multiplication (pre-multiplication) by an elementary matrix represents elementary row operations, while right multiplication (post-multiplication) represents elementary column operations.

Remark 11.32.2 The inverse of elementary matrices has the same format as the orginal ones.

11.33 Permutation Matrices

Remark 11.33.1 When a permutation matrix P is multiplied with a matrix M from the left it will permute the rows of M, when P is multiplied with M from the right it will permute the columns of M.

Remark 11.33.2 The inverse of a permutation matrix is its transpose.

[10] Actually, it should work for all dot product (not necessarily the inner product, which is in the context of Euclidean spaces.)

11.34 Symmetric Matrices

11.35 Projection Matrices

Remark 11.35.1 $P = A(A^TA)^{-1}A^T$, $P = \frac{aa^T}{\|a\|}$

Remark 11.35.2 $P^2 = P$

Remark 11.35.3 Only two eigenvalues possible: 0 and 1. The corresponding eigenvectors form the kernel and range of A, respectively.

Remark 11.35.4 Projection is invertible.

11.36 Normal Matrix

Definition 11.36.1 A *normal matrix* is a square matrix which satisfies

$$\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T \tag{11.27}$$

11.37 Orthogonal Matrices

Definition 11.37.1 An *orthogonal matrix* (*unitary* for a complex matrix) is a normal matrix which further satisfies

$$\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I} \tag{11.28}$$

Or, alternatively,

Remark 11.37.1 $Q^TQ = I$ even if Q is rectangular (but then left-inverse).

Remark 11.37.2 Any permutation matrix P is an orthogonal matrix.

Remark 11.37.3 Orthogonal matrices can be categorized into either the reflection matrix $Ref(\theta)$ which has determinant 1, or the rotation matrix $Rot(\theta)$, which has determinant -1.

Remark 11.37.4 Geometrically, an orthogonal Q is the product of a rotation and a reflection.

Remark 11.37.5 Orthogonal matrix is invariant to 2-norm, that is, suppose Q is an orthogonal matrix, and x a vector, then

$$\|Qx\| = \|x\| \tag{11.29}$$

Remark 11.37.6 Projection matrices are usually not orthogonal, since they are not invariant to 2-norm.

Remark 11.37.7 As a linear transformation, an orthogonal matrix preserves the dot product of vectors (therefore also norm and angle), and therefore acts as an isometry of Euclidean space, such as a rotation or reflection. In other words, it is a unitary transformation.

Remark 11.37.8 The product of two rotation matrices is a rotation matrix, and the product of two reflection matrices is also a rotation matrix. See figure **??**.

11.38 Positive Definite Matrices

Definition 11.38.1 Let $\mathbf{A}$ be an $n \times n$ square matrix. $\mathbf{A}$ is said to be positive definite if

$$\mathbf{x}^T \mathbf{A}\mathbf{x} > 0, \quad \forall \mathbf{x} \neq \mathbf{0} \tag{11.30}$$

Proof The diagonal elements are positive because $a_{kk} = \mathbf{e}_k^T \mathbf{A}\mathbf{e}_k > 0$. The eigenvalues of an s.p.d. matrix are all positive is easy to prove by observing that

$$0 < \mathbf{x}^T \mathbf{A}\mathbf{x} = \mathbf{x}^T \lambda \mathbf{x} = \lambda \|\mathbf{x}\|_2^2$$

The positivity of determinant can be shown by looking at the LDU decomposition. Finally, it is nonsingular because the determinant is nonzero.

Definition 11.38.2 Let $\mathbf{A}$ be an $n \times n$ square matrix. A principal submatrix of $\mathbf{A}$ is obtained by selecting some rows and columns with the *same* index subset of $\{1, \cdots, n\}$.

Definition 11.38.3 Let $\mathbf{A}$ be an $n \times n$ square matrix. A *leading* principal submatrix of $\mathbf{A}$ is a principal submatrix of $\mathbf{A}$ with the index subset $\{1, \cdots, m\}$, for some $m \leq n$.

Proof Suppose $\mathbf{A}_p$ of size p is a principle submatrix of $\mathbf{A}$. Since $\mathbf{A}$ is positive definite, for any nonzero vector $\mathbf{x}$ we have $\mathbf{x}^T \mathbf{A}\mathbf{x} > 0$. Remove the corresponding coordinates of $\mathbf{x}$, same as those removed when creating the principle submatrix, and call it $\mathbf{x}_p$. Then the resulting vector $\mathbf{x}_p^T \mathbf{A}_p \mathbf{x}_p = \mathbf{x}^T \mathbf{A}\mathbf{x} > 0$.

11.39 Numerical Linear Algebra Algorithms

11.40 Matrix Inverse: Binomial inverse theorem, Schur Complement, Blockwise Inversion

Remark 11.40.1 $\mathbf{A}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

Usually, $|\mathbf{AB}| \neq |\mathbf{BA}|$. For example

Example 11.40.1

$$\left|\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix}\right| = |1| = 1$$

$$\left|\begin{bmatrix}0\\1\end{bmatrix}\begin{bmatrix}1 & 1\end{bmatrix}\right| = \left|\begin{bmatrix}0 & 0\\1 & 1\end{bmatrix}\right| = 0$$

However, the **Sylvester's Determinant Theorem** says, as long as **AB** and **BA** are both square matrices,

$$|\mathbf{I} + \mathbf{AB}| = |\mathbf{I} + \mathbf{BA}| \tag{11.31}$$

It is also not true in general that

$$\left|\begin{bmatrix}\mathbf{A} & \mathbf{B}\\\mathbf{C} & \mathbf{D}\end{bmatrix}\right| = |\mathbf{AD} - \mathbf{BC}|$$

unless **C** and **D** are commutable, i.e., $\mathbf{CD} = \mathbf{DC}$. The general formula for block determinant is

$$\left|\begin{bmatrix}\mathbf{A} & \mathbf{B}\\\mathbf{C} & \mathbf{D}\end{bmatrix}\right| = |\mathbf{A}|\,|\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B}| \tag{11.32}$$

which is based on Schur complement.

11.41 The Ax = b Problem

11.42 Solving a Linear System of Equations

Theorem 11.42.1 If $A = \begin{bmatrix}a & b\\c & d\end{bmatrix}$, then A is invertible if $ad - bc \neq 0$, in which case

$$A^{-1} = \frac{1}{ad - bc}\begin{bmatrix}d & -b\\-c & a\end{bmatrix}$$

Remark 11.42.1 Three ways to solve a system of linear equations: by elimination, by determinants (Cramer's Rule???), or by matrix decomposition.

Remark 11.42.2 We prefer to use matrix decomposition to solve a linear system because

1. It takes $\mathcal{O}(n^3)$ to factorize, but once done it can be used to solve systems with different **b** (right hand side).

2. It is numerically more stable than computing $\mathbf{A}^{-1}\mathbf{b}$.
3. For a sparse matrix, the inverse may be dense and may hard to store in memory. Decomposition can overcome this problem.

Remark 11.42.3 Cofactors and Minors. Laplace's Theorem.

Remark 11.42.4 The computation of elimination is $\mathcal{O}(n^3)$, but can be (non-trivially) reduced to $\mathcal{O}(n^{\log_2 7})$.

11.43 The Vector Spaces of a Matrix

Remark 11.43.1 Ax is a combination of the *columns* of A. $b^T A$ is a combination of the *rows* of A. Row picture can be seen as interchapter of (hyper-)planes. Column picture can be seen as combination of columns.

Remark 11.43.2 There are three different ways to look at matrix multiplication:

1. Each entry of AB is the product of a row (of A) and a column (of B)
2. Each *column* of AB is the product of a matrix (of A) and a column (of B)
3. Each *row* of AB is the product of a row (of A) and a matrix (of B)

Remark 11.43.3 Column space is perpendicular to the left null space. Row space is perpendicular to the null space.

11.44 The $\mathbf{Ax} = \lambda\mathbf{x}$ Problem

11.45 Matrix Decomposition

11.46 Decomposition related to solving $\mathbf{Ax} = \mathbf{b}$

11.46.1 LU Decomposition: Schur Complement

Usually, $|\mathbf{AB}| \neq |\mathbf{BA}|$. For example

Example 11.46.1

$$\left|\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix}\right| = |1| = 1$$

$$\left|\begin{bmatrix}0\\1\end{bmatrix}\begin{bmatrix}1 & 1\end{bmatrix}\right| = \left|\begin{bmatrix}0 & 0\\1 & 1\end{bmatrix}\right| = 0$$

However, the **Sylvester's Determinant Theorem** says, as long as $\mathbf{AB}$ and $\mathbf{BA}$ are both square matrices,

$$\left|\mathbf{I}+\mathbf{AB}\right| = \left|\mathbf{I}+\mathbf{BA}\right| \tag{11.33}$$

It is also not true in general that

$$\left|\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}\right| = \left|\mathbf{AD}-\mathbf{BC}\right|$$

unless $\mathbf{C}$ and $\mathbf{D}$ are commutable, i.e., $\mathbf{CD}=\mathbf{DC}$. The general formula for block determinant is

$$\left|\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}\right| = \left|\mathbf{A}\right|\left|\mathbf{D}-\mathbf{CA}^{-1}\mathbf{B}\right| \tag{11.34}$$

which is based on Schur complement.

Now suppose we have a homogeneous linear system

$$\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix} \tag{11.35}$$

To solve for $\mathbf{y}$, if $\mathbf{A}$ is nonsingular, we may multiply the first row by $-\mathbf{CA}^{-1}$ and add to the second, and obtain

$$\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{CA}^{-1} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix} = \begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{0} & \mathbf{D}-\mathbf{CA}^{-1}\mathbf{B}\end{bmatrix} \tag{11.36}$$

Definition 11.46.1 Suppose $\mathbf{M}$ is a square matrix

$$\mathbf{M} = \begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}$$

and $\mathbf{A}$ nonsingular. We denote [11]

$$\mathbf{M}/\mathbf{A} = \mathbf{D}-\mathbf{CA}^{-1}\mathbf{B} \tag{11.37}$$

and call it *the Schur complement of* $\mathbf{A}$ *in* $\mathbf{M}$, or *the Schur complement of* $\mathbf{M}$ *relative to* $\mathbf{A}$.

Theorem 11.46.1

$$det(\mathbf{M}) = det(\mathbf{M}/\mathbf{A})\cdot det(\mathbf{A}) \tag{11.38}$$

$$rank(\mathbf{M}) = rank(\mathbf{M}/\mathbf{A}) + rank(\mathbf{A}) \tag{11.39}$$

[11] It is easy to remember if you multiply the submatrices clockwise.

Remark 11.46.1 For a non-homogeneous system of linear equations

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}$$

We may use Schur complements to write the solution as

$$\mathbf{x} = (\mathbf{M}/\mathbf{D})^{-1}(\mathbf{u} - \mathbf{B}\mathbf{D}^{-1}\mathbf{v}) \tag{11.40}$$

$$\mathbf{y} = (\mathbf{M}/\mathbf{A})^{-1}(\mathbf{v} - \mathbf{C}\mathbf{A}^{-1}\mathbf{u}) \tag{11.41}$$

Theorem 11.46.2 If $\mathbf{M}$ is a positive-definite symmetric matrix, then so is the Schur complement of $\mathbf{D}$ in $\mathbf{M}$.

11.46.2 LDU Decomposition

The LDU decomposition can be viewed as the matrix form of Gaussian elimination. It is used to find the inverse of a matrix, **or computing the determinant of a matrix**.

Remark 11.46.2 The triangular factorization can be written $A = LDU$, where L and U have $1's$ on the diagonal and D is the diagonal matrix of pivots.

11.46.3 Rank Decomposition

11.46.4 Cholesky Decomposition

Definition 11.46.2 The Cholesky decomposition of an s.p.d. matrix $\mathbf{A}$ is of the form

$$\mathbf{A} = \mathbf{L}\mathbf{L}^* \tag{11.42}$$

where $\mathbf{L}$ is a lower triangular matrix, with *real and positive diagonal elements*.

Definition 11.46.3 The Cholesky decomposition of a s.p.d. matrix $\mathbf{A}$ is of the form

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T \tag{11.43}$$

where $\mathbf{L}$ is a lower triangular matrix, with *real and positive diagonal elements*.

Cholesky decomposition is unique. If $\mathbf{A}$ is symmetric semi-positive definite, it still has a decomposition of the form $\mathbf{A} = \mathbf{L}\mathbf{L}^*$, although may not be unique, if the diagonal entries of $\mathbf{L}$ are allowed to be zero. A closely related variant of the classical Cholesky decomposition is the LDL^T *decomposition*:

$$\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^T = (\mathbf{L}\mathbf{D}^{\frac{1}{2}})(\mathbf{D}^{\frac{1}{2}}{}^T\mathbf{L}^T) = (\mathbf{L}\mathbf{D}^{\frac{1}{2}})(\mathbf{L}\mathbf{D}^{\frac{1}{2}})^T \tag{11.44}$$

where the diagonal entries of $\mathbf{L}$ are all ones.

11.46.5 QR Decomposition: Givens Rotation, Householder Transformation

Any square matrix $\mathbf{A}$ may be decomposed as

$$\mathbf{A} = \mathbf{QR} \tag{11.45}$$

where $\mathbf{Q}$ is an orthogonal matrix, and $\mathbf{R}$ an upper triangular matrix. This is called the QR decomposition. It is essentially a change of basis process, and can be obtained by using the Gram-Schmidt process.

11.47 Decomposition related to solving $\mathbf{Ax} = \lambda\mathbf{x}$

11.47.1 Eigendecomposition

Suppose a square matrix $\mathbf{M}$ of order n is diagonalizable, i.e., it has n linearly independent eigenvectors, then since

$$\mathbf{MQ} = \mathbf{Q\Lambda} \tag{11.46}$$

where the columns of $\mathbf{Q}$ are eigenvectors of $\mathbf{M}$ (hence invertible), and $\mathbf{\Lambda}$ is a diagonal matrix with eigenvalues of $\mathbf{M}$ as entries. Then we have

$$\mathbf{M} = \mathbf{Q\Lambda Q}^{-1} \tag{11.47}$$

11.47.2 Jordan Decomposition

11.47.3 Schur Decomposition

11.47.4 Singular Value Decomposition (SVD)

11.47.5 QZ Decomposition

11.48 Other Decompositions

11.48.1 Polar Decomposition

11.49 Minors and Cofactors

11.50 Definition

Definition 11.50.1 General definition of a minor.

Let $\mathbf{A}$ be an $m \times n$ matrix and k an integer with $0 < k \leq \min m, n$. A $k \times k$ minor of $\mathbf{A}$ is the determinant of a $k \times k$ matrix obtained from $\mathbf{A}$ by deleting $m - k$ rows and $n - k$ columns. For such a matrix there are a total of $\binom{m}{k} \cdot \binom{n}{k}$ minors of size $k \times k$.

Definition 11.50.2 First minors and cofactors.

If A is a square matrix, then the minor of the entry in the i-th row and j-th column (also called the (i, j) minor, or a first minor, is the determinant of the submatrix formed by deleting the i-th row and j-th column. This number is often denoted M_{ij}. The (i, j) cofactor is obtained by multiplying the minor by $(-1)^{i+j}$.

Example 11.50.1 To illustrate these definitions, consider the following 3 by 3 matrix,

$$\begin{bmatrix} 1 & 4 & 7 \\ 3 & 0 & 5 \\ -1 & 9 & 11 \end{bmatrix} \tag{11.48}$$

To compute the minor M_{23} and the cofactor C_{23}, we find the determinant of the above matrix with row 2 and column 3 removed.

$$M_{2,3} = \det \begin{bmatrix} 1 & 4 & \square \\ \square & \square & \square \\ -1 & 9 & \square \end{bmatrix} = \det \begin{bmatrix} 1 & 4 \\ -1 & 9 \end{bmatrix} = (9 - (-4)) = 13$$

So the cofactor of the (2,3) entry is $C_{23} = (-1)^{2+3}(M_{23}) = -13$.

An important application of cofactors is the **Laplace's formula** for the expansion of determinants.

$$\det(\mathbf{A}) = \sum_{i=1}^{n} a_{ij} C_{ij} = \sum_{j=1}^{n} a_{ij} C_{ij} \tag{11.49}$$

If $k \neq i$, we see that

$$\sum_{j=1}^{n} a_{kj} C_{ij} = 0 \tag{11.50}$$

Similarly, if $k \neq j$

$$\sum_{i=1}^{n} a_{ik} C_{ij} = 0 \tag{11.51}$$

This is essentially the determinant of a matrix with the k-th row the same as the i-th row, or the k-th column the same as the j-th column, which is zero.

11.51 The Cramer's Rule and the Adjugate Matrix

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &=& b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &=& b_2 \\ \vdots & \vdots & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &=& b_n \end{array} \tag{11.52}$$

If we multiply the above by the row vector of cofactors of the 1^{st} column, $[C_{11}, C_{21}, \cdots, C_{n1}]$, we obtain

$$[\det(\mathbf{A}), 0, \cdots, 0] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = [C_{11}, C_{21}, \cdots, C_{n1}]\mathbf{b} \tag{11.53}$$

The left hand side used Equation 11.51. The right hand side is nothing but the determinant of a matrix with the first column replaced by $\mathbf{b}$.

Similarly, we can multiply the linear system by the row vector of cofactors of the $2^{nd}, 3^{rd}, \cdots, n^{th}$, and we obtain

$$\det(\mathbf{A})\mathbf{x} = \begin{bmatrix} C_{11} & \cdots & C_{n1} \\ C_{12} & \cdots & C_{n2} \\ \vdots & & \vdots \\ C_{1n} & \cdots & C_{nn} \end{bmatrix} \mathbf{b} \tag{11.54}$$

which gives us

$$\det(\mathbf{A}) = \begin{bmatrix} C_{11} & \cdots & C_{1n} \\ C_{21} & \cdots & C_{2n} \\ \vdots & & \vdots \\ C_{n1} & \cdots & C_{nn} \end{bmatrix}^T \mathbf{A} \tag{11.55}$$

The matrix on the right

$$\text{adj}(\mathbf{A}) = \mathbf{C}^T = \begin{bmatrix} C_{11} & \cdots & C_{1n} \\ C_{21} & \cdots & C_{2n} \\ \vdots & & \vdots \\ C_{n1} & \cdots & C_{nn} \end{bmatrix}^T \tag{11.56}$$

is called the adjugate matrix of $\mathbf{A}$, which is the transpose of the cofactor matrix $\mathbf{C}$.

11.52 Integers and Equivalence Relations

Theorem 11.52.1 Well Ordering Principle. Every nonempty set of positive integers contains a smallest member.

Theorem 11.52.2 Division Algorithm. Let a and b be integers with $b > 0$. Then there exist unique integers q and r with the property that $a = bq + r$, where $0 \leq r < b$. (Note: a and q could be negative.)

Theorem 11.52.3 GCD (Greatest Common Divisor) is a Linear Combination. For any nonzero integers a and b, there exist integers s and t such that $gcd(a, b) = as + bt$. Moreover, $gcd(a, d)$ is the smallest positive integer of the form $as + bt$.

Corollary 11.52.1 If a and b are relatively prime, then there exist integers s and t such that $as + bt = 1$.

Theorem 11.52.4 If $a \mod n = a'$ and $b \mod n = b'$, then $(a+b) \mod n = (a' + b') \mod n$ and $(ab) \mod n = (a'b') \mod n$.

Part III

Statistics and Machine Learning

11.53 Statistics Preface

This booklet is divided into 7 Chapters. The first chapter introduces the definitions of basic concepts, such as event, sample space, and probability space. Followed in the next chapter, we will discuss the relationship between two or more events when they interplay with each other. The third chapter formally brings in random variables and vectors, as a basis to develop their quantitative measure and characteristic functions later in chapter four. Chapter five includes some well-known limit theorems, which is useful for asymptotic analysis. The last two chapters will discuss several selected topics in probability theory, and provide a summary of common distributions.

Six types of statistical analysis:

1. Descriptive
2. Exploratory
3. Inferential (parameters)
4. Predictive (what will happen)
5. Causal (why it happens)
6. Mechanistic (how to deal with it)

A correlation matrix is defined as:

$$\text{Corr}(\mathbf{X}) = (\text{diag}(\mathbf{\Sigma}))^{-\frac{1}{2}}\mathbf{\Sigma}(\text{diag}(\mathbf{\Sigma}))^{-\frac{1}{2}} \tag{11.57}$$

Both covariance matrices and correlation matrices are symmetric semipositive definite matrices.

Main References:

1. Extending the Linear Model with R, by Julian J. Faraway
2. Categorical Data Analysis 3rd Edition, by Alan Agresti
3. Generalized, Linear, and Mixed Models, by Charles E. McCulloch, Shayle R. Searle, John M. Neuhaus
4. An Introduction to Generalized Linear Models, by Annette J. Dobson, Adrian Barnett
5. Generalized Linear Models, by P. McCullagh, John A. Nelder

11.54 Collecting Data: Experiments and Surveys

11.55 Design of Experiments

11.56 Statistical Survey

11.57 Opinion Poll

11.58 Sampling

11.58.1 Sampling Distribution

11.58.2 Sampling: Stratified Sampling, Quota Sampling

11.58.3 Biased Sample: Spectrum Bias, Survivorship Bias

11.59 Describing Data

11.60 Average: Mean, Median, and Mode

11.61 Measures of Scale: Variance, Standard Deviation, Geometric Standard Deviation, and Median Absolute Deviation

11.62 Correlation and Dependence

11.63 Outlier

11.64 Statistical Graphics: Histogram, Frequency Distribution, Quantile, Survival Function, and Failure Rate

11.65 Filtering Data

11.66 Recursive Bayesian Estimation

11.66.1 Kalman Filter

11.66.2 Particle Filter

11.67 Moving Average

11.68 Linear Regression Models

11.69 Introduction

Generalized linear models include as special cases, linear regression and analysis-of-variance models, logit and probit models for quantal responses, log linear

models and multinomial response models for counts and some commonly used models for survival data.

The second-order properties of the parameter estimates are insensitive to the assumed distributional form: the second-order properties depend mainly on the assumed variance-to-mean relationship and on uncorrelatedness or independence.

Data types:

11.70 Simple Linear Regression

11.70.1 Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{11.58}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

11.70.2 Estimated Regression Function

$$b_1 = \rho_{XY} \cdot \frac{s_Y}{s_X} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sum_{i=1}^n \left[\frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] Y_i \tag{11.59}$$

$$b_0 = \bar{Y} - b_1 \bar{X} \tag{11.60}$$

$$\hat{\sigma}^2 = \frac{MSE}{n-2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2} \tag{11.61}$$

Notice, $\sum (X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2$, and $b_1 = \rho \cdot \frac{s_Y}{s_X}$, where ρ is the correlation between X and Y and s_Y, s_X are standard error of Y and X, respectively.

The slope of the fitted line is equal to the correlation between y and x corrected by the ratio of standard deviations of these variables. The intercept of the fitted line is such that it passes through the center of mass $(\bar{x}, \bar{y})$ of the data points.

Another way of writing the estimated regression function is

$$\hat{Y}_i = \bar{Y} + b_1 (X_i - \bar{X}) \tag{11.62}$$

Notice, $\bar{Y}$ and b_1 are uncorrelated (check it using the fact that $b_1 = \sum_{i=1}^n k_i Y_i$).

11.70.3 Inference About b_1 and b_0

Since $SSE/\sigma^2 \sim \chi^2_{n-2}$, and $\frac{s^2\{b_1\}}{\sigma^2\{b_1\}} \sim \frac{\chi^2_{n-2}}{n-2}$

$$\frac{b_1 - \beta_1}{s\{b_1\}} = \frac{b_1 - \beta_1}{\sigma\{b_1\}} \Big/ \frac{s\{b_1\}}{\sigma\{b_1\}} \sim \frac{z}{\sqrt{\frac{\chi^2_{n-2}}{n-2}}} = t_{n-2} \tag{11.63}$$

so the confidence interval for b_1, with confidence level α is

$$b_1 \pm t(1-\alpha/2; n-2)s\{b_1\} \tag{11.64}$$

or

$$b_1 \mp t(\alpha/2; n-2)s\{b_1\} \tag{11.65}$$

Similarly, the confidence interval for b_0, with confidence level α is

$$b_0 \pm t(1-\alpha/2; n-2)s\{b_0\} \tag{11.66}$$

or

$$b_0 \mp t(\alpha/2; n-2)s\{b_0\} \tag{11.67}$$

The power of testing $\beta_1 = \beta^{H_0}$ is $Power = P\{|t^*| > t(1-\alpha/2; n-2)|\delta\}$, where $\delta = \frac{|\beta_1 - \beta^{H_0}|}{\sigma\{b_1\}}$. Similar for β_0.

11.70.4 Properties of k_i

$$k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{11.68}$$

$$\sum_{i=1}^{n} k_i = 0 \tag{11.69}$$

$$\sum_{i=1}^{n} k_i X_i = 1 \tag{11.70}$$

$$\sum_{i=1}^{n} k_i^2 = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{11.71}$$

The second and third identities hold as a requirement for the unbiasness, since

$$E(b_1) = E\left(\sum k_i Y_i\right) = E\left(\sum k_i(\beta_0 + \beta_1 X_i)\right) = E\left(k_i \sum \beta_0 + \beta_1 \sum k_i X_i\right) = \beta_1$$

requires $\sum k_i = 0$ and $\sum X_i k_i = 1$. The fourth identity ensures the attainment of the minimum variance.

11.70.5 Properties of e_i

$$e_i = Y_i - \hat{Y}_i \tag{11.72}$$

$$\sum e_i = 0 \tag{11.73}$$

$$\sum X_i e_i = 0 \tag{11.74}$$

$$\sum \hat{Y}_i e_i = 0 \tag{11.75}$$

11.70.6 Properties of b_1 and b_0

$$b_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum(X_i - \bar{X})^2}\right) \tag{11.76}$$

$$b_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{X}^2}{\sum(X_i - \bar{X})^2}\right) \tag{11.77}$$

where σ^2 can be estimated by the MSE, i.e., $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$

Now, since

$$b_1 = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sum_{i=1}^{n}(X_i - \bar{X})Y_i \tag{11.78}$$

$$= \sum_{i=1}^{n} k_i Y_i \tag{11.79}$$

where $k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$, we have

$$\sum k_i = 0 \tag{11.80}$$

$$\sum X_i k_i = 1 \tag{11.81}$$

$$\sum k_i^2 = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{11.82}$$

The first two identity hold as a requirement for the unbiasness, since

$$E(b_1) = E\left(\sum k_i Y_i\right) = E\left(\sum k_i(\beta_0 + \beta_1 X_i)\right) = E\left(\beta_0 \sum k_i \beta_0 + \beta_1 \sum k_i X_i\right) = \beta_1$$

requires $\sum k_i = 0$ and $\sum X_i k_i = 1$. The third identity ensures the attainment of the minimum variance.

	Estimate	Expectation	Variance
Y_i	$\hat{Y}_i$	$\beta_0 + \beta_1 X_i$	σ^2
b_1	$\frac{\sum(X_i-\bar{X})Y_i}{\sum(X_i-\bar{X})^2}$	β_1	$\sigma^2 \cdot \frac{1}{\sum(X_i-\bar{X})^2}$
b_0	$\bar{Y} - b_1\bar{X}$	β_0	$\sigma^2 \cdot \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i-\bar{X})^2}\right]$
$\hat{Y}_h$	$\bar{Y} + b_1(X_h - \bar{X})$	$\beta_0 + \beta_1 X_h$	$\sigma^2 \cdot \left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]$
$\hat{Y}_{h(new)}$	$\bar{Y} + b_1(X_h - \bar{X})$	$\beta_0 + \beta_1 X_h$	$\sigma^2 \cdot \left[1 + \frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]$
$\hat{Y}_{h(new_m)}$	$\bar{Y} + b_1(X_h - \bar{X})$	$\beta_0 + \beta_1 X_h$	$\sigma^2 \cdot \left[\frac{1}{m} + \frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]$
e_i	$Y_i - \hat{Y}_i$	0	$1 - h_{ii}$

Table 11.1: Simple Linear Regression

In particular, when $X_h = 0$ we obtain the formulas for b_0, and when $X_h - \bar{X} = 1$ we obtain the formulas for b_1.

11.70.7 ANOVA of Simple Linear Regression Model

$$SSTO = SSR + SSE \tag{11.83}$$

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \sum_{i=1}^{n}(\hat{Y}_i - \bar{y}) + \sum_{i=1}^{n}(\bar{y} - \hat{Y}_i) \tag{11.84}$$

SSR can also be computed as $SSR = b_1^2 \sum_{i=1}^{n}(X_i - \bar{X})$, so given the same "distribution" of X, the steeper the slope of the regression line, the higher the SSR, and hence the better fit of the model.

To test $H_0 : \beta_1 = 0$, we use $F = \frac{SSR}{SSE}$. There is equivalence between an F test and a t test: $[t(1-\alpha/2, n-2)]^2 = F(1-\alpha, n-2)$.

11.71 Generalized Linear Regression Models

11.72 Survival Analysis

$$S(t) = \exp\left[-\int_0^t \lambda(u)du\right] \tag{11.85}$$

$$L(\lambda) = \prod_{i=1}^{n} [\lambda(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} \tag{11.86}$$

where $S(t)$ is the survival function, and $\lambda(t)$ is the hazard function.

	Estimate	Standard Error	NOTE
S	$\hat{S(t)} = \prod \frac{n_j - d_j}{n_j}$	$\hat{S(t)}\sqrt{\sum \frac{d_i}{n_j(n_j-d_j)}}$	
Λ	$-\log \hat{S(t)}$	$\sqrt{\sum \frac{d_i}{n_j(n_j-d_j)}}$	
λ	$\frac{\sum \delta_i}{\sum (X_i - V_i)}$	$\frac{\hat{\lambda}}{\sqrt{\sum \delta_i}}$	

Table 11.2: Survival Analysis

11.73 Analysis of Variance (ANOVA)

11.74 Multivariate Analysis

11.75 Principal Component Analysis (PCA)

ALgebraically, principal components are particular linear combinations of the p random variables $X_1, \cdots, X_p$. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $X_1, \cdots, X_p$ as the coordinate axes.

11.76 Factor Analysis

11.77 Cluster Analysis

11.78 Discriminant Analysis

11.79 Correspondence Analysis

11.80 Canonical Correlation Analysis (CCA)

11.81 Multidimensional Scaling (MDS)

11.82 Modeling Sample Data

11.83 Density Estimation

11.83.1 Kernel Density Estimation

11.83.2 Multivariate Kernel Density Estimation

11.84 Time Series

11.85 Robust Statistics

11.86 Modeling Population Data: Statistical Inference

11.87 Bayesian Inference

11.87.1 Bayes' theorem, Bayes Estimator, Prior Distribution, Posterior Distribution, Conjugate Prior, and All That

11.88 Frequentist Inference

11.88.1 Statistical Hypothesis Testing: Null, Alternative, P-value, Significance level, power, likelihood-ratio test, goodness-of-fit, confidence interval, M-estimator, Trimmed Estimator

11.89 Non-parametric Statistics

11.89.1 Nonparametric Regression, Kernel Methods

11.90 Making Decisions: Decision Theory

11.91 Optimal Decision, Type I and Type II errors

11.92 Correlation and Causation

When a statistical test shows a correlation between A and B, there are usually

1. A causes B.

2. B causes A.

3. A and B both partly cause each other.

4. A and B are both caused by a third factor, C.

5. The observed correlation was due purely to chance.

11.93 Theory of Linear Models

11.94 Linear Models, Estimable Functions, Least Squares Estimates=LSE, Normal Equations, Projections, Gauss Markov theorem, BLUE

11.95 Multivariate Normal Distribution and Distribution of Linear and Quadratic Forms

11.96 Properties of LSE and Generalized LSE

11.97 General Linear Hypothesis=GLH, Testing of GLH

11.98 Orthogonalization of Design Matrix and Canonical Reduction of GLH; Adding Variables To The Model

11.99 Correlation, Multiple Correlation and Partial Correlation

11.100 Confidence Regions and Prediction Regions

11.101 Simultaneous Confidence Sets, Bonferroni, Scheffe Projection Method, Tukey Studentized Range

11.102 Introduction to Design of Experiments, ANOVA and ANOCOVA, Factorial and Block Designs, Random, Fixed and Mixed Models, Components of Variance

11.103 Hierarchial Bayes Analysis of Variance; (Schervish Ch. 8, 8.1,8.2) Partial Exchangeability and Hierarchical, Models, Examples and Representations, Normal One Way ANOVA and Two Way Mixed Model ANOVA

11.104 Mathematical Statistics

11.105 Degrees of Freedom

$$E(y) = b'(\theta) \tag{11.88}$$
$$Var(y) = b''(\theta)a(\phi) \tag{11.89}$$

11.109 Cramer-Rao Theorem

11.110 Data, Models, Statistics, Parameters

11.111 Distributions of Functions of a Random Variable

Theorem 11.111.1 From Casella & Berger theorem 2.1.5) Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is a monotone function. Suppose that $f_X(x)$ is continuous and that $g^{-1}(y)$ has a continuous derivative. Then the pdf of Y is given by

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right| \tag{11.90}$$

11.112 Decision Theory (Bayes and Minimax Criteria, Risk Functions, Estimation and Testing in Terms of the Decision Theoretic Framework

11.113 Bayesian Models, Conjugate (and Other) Prior Distributions

11.114 Prediction (Optimal MSPE and Optimal Linear MSPE)

11.115 Sufficiency (Factorization theorem)

11.116 Natural Sufficient Statistics

11.117 Minimal Sufficiency

11.118 Estimation (Least Squares, MLE, Frequency Plug-in, Method of Moments, Combinations of These)

11.119 Exponential Families & Properties, Canonical Exponential Families (& Fisher Information)

11.120 Information Inequality, Fisher Information, UMVU Estimates, Cramer-Rao Lower Bound

11.121 Neyman-Pearson Testing Theory (Form, MP Test, UMP Test, MLR Family, Likelihood Ratio Tests)

11.122 Asymptotic Approximation / Large Sample Theory (Consistency, Delta Method, Asymptotic Normality of MLE, Slutsky's theorem, Efficiency, Pearson's Chi-Square)

11.123 Selected Topics

11.124 M-estimator

11.125 Sweep Operator

11.126 Information Geometry

11.127 Boostrap

Followed in the next chapter, we will discuss the relationship between two or more events when they interplay with each other. The third chapter formally brings in random variables and vectors, as a basis to develop their quantitative measure and characteristic functions later in chapter four. Chapter five includes some well-known limit theorems, which is useful for asymptotic analysis. The last two chapters will discuss several selected topics in probability theory, and provide a summary of common distributions.

11.129 Elementary Theory of Probability

11.130 Combinatorial Analysis

11.131 Axioms

There are two important rules in combinatorics: the rule of sum, and the rule of product.

The rule of sum says, if we have a ways to finish a task using one method and alternatively, b ways to finish the same task using another method, then there are ab ways of finish this task. More generally,

$$|S_1 \cup S_2 \cup \ldots \cup S_n| = |S_1| + |S_2| + \ldots + |S_n| \tag{11.91}$$

One extension of the rule of sum is the inclusion-exclusion principle, which does not require sets A_i to be disjoint. This does include the rule of sum, in that if sets A_i are disjoint, the terms from the second to the last are all zero.

$$\begin{aligned} |\bigcup_{i=1}^{n} A_i| = &\sum_{i=1}^{n} |A_i| - \sum_{1\le i<j\le n} |A_i \cap A_j| + \sum_{1\le i<j<k\le n} |A_i \cap A_j \cap A_k| - \cdots \\ &+ (-1)^{n-1} |A_1 \cap \cdots \cap A_n| \end{aligned} \tag{11.92}$$

The rule of product says, if finishing one task requires two steps, and there are a ways to choose in the first step and b ways to choose in the second step, then there are ab ways to finish this task. More generally,

$$|S_1 \times S_2 \times \cdots \times S_n| = |S_1| \cdot |S_2| \cdots |S_n| \tag{11.93}$$

11.132 Binomial Coefficient and Its Applications

11.132.1 Binomial Coefficient

We list here some of the useful binomial identities, all numbers are nature number (not including 0).

$$\binom{n}{k} = \binom{n}{n-k} \tag{11.94}$$

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n \tag{11.95}$$

$$\binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1} \tag{11.96}$$

From the famous Pascal's rule,

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1} \tag{11.97}$$

There is another form which is equivalent to equation 11.97,

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} \tag{11.98}$$

Here is an example that uses the *logarithmic differentiation*, $f' = f\dot{[}\ln(f)]'$.

$$\frac{d}{dt}\binom{t}{k} = \binom{t}{k}\sum_{i=0}^{k-1}\frac{1}{t-i} \tag{11.99}$$

A list of series that involves binomial coefficients,

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n \tag{11.100}$$

$$\sum_{k=0}^{n} k\binom{n}{k} = n2^{n-1} \tag{11.101}$$

$$\sum_{k=0}^{n} k^2\binom{n}{k} = (n+n^2)2^{n-2} \tag{11.102}$$

These can all be obtained by examining the function value or derivatives of the function $(1+x)^\alpha$, where α could be any real number, and $|x| < 1$.

There are some identities that could be proved using combinatorial analysis, such as *double counting*. Here is an example,

$$\sum_{k=1}^{n} \binom{n}{k}\binom{k}{q} = 2^{n-q}\binom{n}{q} \tag{11.103}$$

The left side of equation 11.103 counts the number of ways of selecting k elements first, and then choosing q elements from the resulting subset. These q elements could be identical for different k. The right hand side of the equation says this is equivalent to first choosing q elements directly from the set, and merging them into one of the 2^{n-q} subset of the set containing all but those selected q elements.

Another example is,

$$\sum_{m_1=0}^{n_1} \binom{n_1}{m_1}\binom{n_2}{m-m_1} = \binom{n}{m} \tag{11.104}$$

This simply means choosing $m = m_1 + m_2$ objects from a set of $n = n_1 + n_2$ objects is equivalent to choosing m_1 objects from n_1 objects, and m_2 objects from n_2 objects.

Sometimes, knowing the bounds and asymptotic formulae could be helpful.

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \frac{n^k}{k!} \leq \left(\frac{n \cdot e}{k}\right)^k \tag{11.105}$$

$$\binom{2n}{n} \sim \frac{4^n}{\sqrt{\pi n}}, \text{ as } n \to \infty \tag{11.106}$$

11.132.2 Bernoulli Distribution

11.132.3 The i.i.d. Case: Binomial Distribution

11.132.4 The Batch Mode Case: Hypergeometric Distribution

11.133 Multinomial Coefficient and Its Applications

11.133.1 Multinomial Coefficient

The notion of *multinomial coefficient* is a generalization of binomial coefficient, which is defined in the multinomial theorem:

$$(x_1 + x_2 + \ldots + x_r)^n = \sum_{(n_1,\ldots,n_r):n_1+\ldots,n_r=n} \binom{n}{n_1, n_2, \ldots, n_r} x_1^{n_1} x_2^{n_2} \ldots x_r^{n_r}$$

We call $\binom{n}{n_1,n_2,\ldots,n_r}$ the multinomial coefficient.

Problem 11.133.1 A set of n distinct items is to be divided into r distinct groups of respective sizes $n_1, n_2, \ldots, n_r$, where $\sum_{i=1}^r n_i = n$. How many different divisions are possible?

Note that there are $\binom{n}{n_1}$ possible choices for the first group; for each choice of the first group there are $\binom{n}{n-n_1}$ possible choices for the second group; and so on. Hence it follows that there are

$$\begin{aligned}
&\binom{n}{n_1}\binom{n-n_1}{n_2}\ldots\binom{n-n_1-n_2-\ldots-n_{r-1}}{n_r} \\
&= \frac{n!}{(n-n_1)!n_1!}\frac{(n-n_1)!}{(n-n_1-n_2)!n_2!}\ldots\frac{(n-n_1-n_2-\ldots-n_{r-1}!}{(0)!n_r!} \\
&= \frac{n!}{n_1!n_2!\ldots n_r!}
\end{aligned}$$

possible divisions.

Alternatively, we can first permute these n items, where there are $n!$ such orderings. The first n_1 elements are assigned to group 1, the next n_2 elements are assigned to group 2, and so on. However, for example, keeping all but n_i group fixed, this method would generate $n_i!$ equivalent divisions (note the order within a group does not matter). Therefore, we need to cancel out the equivalent-group effect by dividing $n_1!n_2!\dots n_r!$. Finally, the multinomial coefficient is,

$$\binom{n}{n_1, n_2, \dots, n_r} = \frac{n!}{n_1!n_2!\dots n_r!} \tag{11.107}$$

11.134 Categorical Distribution

11.135 Multinomial Distribution

11.136 Multiset Coefficient and Its Applications

11.136.1 Multiset Coefficient

The notion of multiset (or bag) is a generalization of the notion of set in which members are allowed to appear more than once.

The number of times an element belongs to the multiset is the *multiplicity* of that member. The total number of elements in a multiset, including repeated memberships, is the *cardinality* of the multiset. For example, in the multiset {a, a, b, b, b, c} the multiplicities of the members a, b, and c are respectively 2, 3, and 1, and the cardinality of the multiset is 6.

The number of multisets of cardinality k, with elements taken from a finite set of cardinality n, is called the *multiset* coefficient or *multiset number*, and is denoted as $\left(\!\binom{n}{k}\!\right)$. It is equivalent to asking, with replacement, the number of all possible combinations of making k draws from a urn with n distinguishable balls labeled $1 \dots n$.

Problem 11.136.1 With replacement, how many possible combinations to make k draws from a urn with n distinguishable balls labeled $1 \dots n$?

Problem 11.136.2 Suppose k balls that are indistinguishable from each other are to be distributed into n distinguishable (non-empty) urns, how many different outcomes are possible?

Now, we are ready to go back to our original prob. prob 11.136.1 could be asked this way: if however, we allow empty urns, how many outcomes are possible?

Another beautiful explanation is to construct an equivalent mapping. Note, the rule of drawing a series of k numbers $a_1, a_2, \dots, a_k$ from the set $\{1, 2, \dots, n\}$ with repetition is

$$1 \leq a_1 \leq a_2 \leq \dots \leq a_k \leq n$$

Now, a new series of k numbers $b_1, b_2, \ldots, b_k$ can be constructed as follows

$$\begin{array}{ccccccc} 1 \leq a_1 < & a_2+1 < & \ldots < & a_k+k-1 \leq n+k-1 \\ \downarrow & \downarrow & & \downarrow \\ b_1 & b_2 & \ldots & b_k \end{array}$$

Note that $b_1 < b_2 < \ldots < b_k$. This is a model without replacement, and is a one-to-one mapping of the original prob. Under the model of drawing k times from $n+k-1$ balls without replacement, the number of all possible combinations is $\binom{n+k-1}{k}$, which is the same as what we obtained earlier.

A last thing worth noting is that, as we mentioned earlier, the number of all multinomial coefficients is also $\binom{k+n-1}{n-1}$.

11.137 Selected Topics

11.137.1 Double Factorial

11.137.2 Stirling Numbers

11.138 The Bertrand's Ballot prob

The Bertrand's ballot prob was first introduced by Joseph Bertrand in 1887, in the form of: "In an election where candidate A receives p votes and candidate B receives q votes with $p > q$, what is the probability that A will be strictly ahead of B throughout the count?"

$$\frac{a}{a+b}\frac{(a-1)-b}{(a-1)+b} + \frac{b}{a+b}\frac{a-(b-1)}{a+(b-1)} = \frac{a-b}{a+b}$$

This proves that the theorem is true for all $p > q \geq 0$.

The number of the unfavorable cases is $2 \times \binom{p+q-1}{q-1}$, of which half starts with "A" and another half starts with "B", as explained above. The number of favorable cases is $\binom{p+q-1}{p-1} - \binom{p+q-1}{q-1}$. Actually, $\frac{\binom{p+q-1}{p-1}-\binom{p+q-1}{q-1}}{\binom{p+q}{p}} = \frac{p-q}{p+q}$.

Consider now the prob to find the probability that the second candidate is never ahead (i.e. ties are allowed); the solution is $\frac{p+1-q}{p+1}$. This is simply seen by awaring the following equivalent description:

- same as the basic version, ties are NOT allowed; but,
- there are $p+1$ votes for candidate A and q votes for candidate B;
- the first vote is for candidate A;

The probability can then be computed as:

$$\begin{aligned} P(\text{A winning } \textit{with} \text{ ties}) &= P(\text{A winning } \textit{without} \text{ ties} \mid \text{the first vote is A}) \\ &= \frac{P(\text{A winning without ties } \textit{and} \text{ the first vote is A})}{P(\text{the first vote is A})} \\ &= \frac{(p+1-q)/(p+1+q)}{(p+1)/(p+1+q)} = \frac{p+1-q}{p+1} \end{aligned}$$

Another way to look at this prob is to model it as the following: represent a voting sequence as a lattice path on the Cartesian plane and,

- Start the path at (0, 0);
- Each time a vote for the first candidate is received move right 1 unit;
- Each time a vote for the second candidate is received move up 1 unit.

Each such path corresponds to a unique sequence of votes and will end at (p, q). A sequence is "good" exactly when the corresponding path never goes above the diagonal line $y = x$; equivalently, a sequence is "bad" exactly when the corresponding path touches the line $y = x + 1$. For each "bad" path P, define a new path P' by reflecting the part of P up to the first point it touches the line across it. P' is a path from (-1, 1) to (p, q). The same operation applied again restores the original P. This produces a one-to-one correspondence between the "bad" paths and the paths from (-1, 1) to (p, q). The number of these paths is $\binom{p+q}{q-1}$. So the probability asked is $\frac{\binom{p+q}{q}-\binom{p+q}{q-1}}{\binom{p+q}{p}} = \frac{p+1-q}{p+1}$.

An interesting application of this is the famous Catalan number formula, which can be introduced under the random walk story. A random walk on the integers is to take n steps of unit length, beginning at the origin and ending at the point m, that never become negative. Assuming n and m have the same parity and $n \geq m \geq 0$, this number is, according to the Bertrand's Ballot prob allowing ties,

$$\binom{n}{\frac{n+m}{2}} - \binom{n}{\frac{n+m}{2}+1} = \frac{m+1}{\frac{n+m}{2}+1}\binom{n}{\frac{n+m}{2}}$$

Here, $p + q = n$ and $p - q = m$, compared to our used settings. When $m = 0$ and n is even, this gives the Catalan number $\frac{1}{\frac{n}{2}+1}\binom{n}{\frac{n}{2}}$.

Let's tweak this prob a little bit more. Let's say candidate A starts at a "bonus" votes, not 0. That is to say, the system goes from $(0, a)$ to $(p + q, p + a - q)$. If we do not allow ties, all paths hit the x-axis will be unfavorable, and the number of these paths equals the number of paths from $(0, a)$ to its "mirror" point $(p + q, -p - a + q)$. So, there are in total $\binom{p+q}{p+a}$ unfavorable paths, and the probability is $1 - \frac{\binom{p+q}{p+a}}{\binom{p+q}{p}}$. Note when $a = 0$, there is already a tie, and the question really should be asked as: "What if the first count is A, and then the process never has a tie". So the probability should compute as:

$$1 - \frac{p}{p+q}\frac{\binom{p-1+q}{p-1}}{\binom{p-1+q}{p-1}} = \frac{p-q}{p+q}$$

It should have no prob with $a > 0$.

If we allow ties, the "mirror" point should be reflected against $y = -1$, so it becomes $(p+q, -2-p-a+q)$. Now, suppose we go up u steps and go down d steps. Solving the following equations:

$$\begin{aligned} u + d &= p + q \\ u + a - d &= -2 - p - a + q \end{aligned}$$

gives us $u = q-a-1$ and $d = p+a+1$. So there are $\binom{p+q}{p+a+1}$ paths unfavorable, and the probability is then $1 - \frac{\binom{p+q}{p+a+1}}{\binom{p+q}{p}}$. When $a = 0$, this becomes $\frac{p+1-q}{p+1}$, which agrees with what we obtained earlier.

In summary, if one starts at $(0, a)$ and ends at $(p+q, p+a-q)$, the number of unfavorable paths is $\binom{p+q}{p+a}$ if not allowing ties, $\binom{p+q}{p+a+1}$ if allowing ties.

11.139 Catalan Number

In chapter 11.138, we first met Catalan number from the generalized Bertrand's Ballot prob. We write here again the definition of Catalan number, with an intuitive interpretation.

The Catalan number is defined as,

$$C_n = \frac{1}{n+1}\binom{2n}{n} \tag{11.108}$$

The underlying story reads: Given two urns, one with n red balls and the other with n black balls, we want to draw one ball at a time (either red or black), such that at no time the number of pre-specified color is less than its alternative.

Since the Catalan number is associated with two equal-sized sets, it is oftentimes co-occurrent with the words "pair", "full binary", and etc.

11.140 Conditional Probability

Definition 11.140.1 If $P(F) > 0$, then

$$P(E|F) = \frac{P(E,F)}{P(F)} \tag{11.109}$$

$P(E|F)$ is called the conditional probability of E given F. Conditional probability agrees with Definition 11.145.3, and should be treated in the same way.

Definition 11.140.2 The multiplication rule

$$P(E_1, E_2, \ldots, E_n) = P(E_1)P(E_2|E_1)\ldots P(E_n|E_1, \ldots, E_{n-1}) \tag{11.110}$$

Definition 11.140.3 Bayes' Formula

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \tag{11.111}$$

where $P(A_i)$ is sometimes called the prior distribution, $P(B|A_i)$ the likelihood function, and $P(A_i|B)$ the posterior distribution. The partition function $P(B)$ can be computed using the law of total probability

$$P(B) = \sum_{j=1}^{n} P(B|A_j)P(A_j) \tag{11.112}$$

Definition 11.140.4 Two events E and F are said to be *independent* if

$$P(E, F) = P(E)P(F) \tag{11.113}$$

A set of events are independent if every finite subset of these events is independent.

11.141 Conditional Probability

11.142 Conditional Expectation

11.143 Conditional Independence

11.144 Probaility Space

11.145 Sample Space, Events, and Probability

Definition 11.145.1 A sample space S is a set of all possible outcomes of an experiment. An outcome is also called a sample point.

Note, when an experiment consists of several repetitions, each one of them is called a *trial*. As an example, if one decides to toss a coin 42 times, we can call each toss a trial of the experiment composed of 42 ones.

Definition 11.145.2 An event E is a subset of the sample space S. If the outcome of the experiment is contained in E, then we say that E occurred.

Definition 11.145.3 In short, a probability space is a measure space such that the measure of the whole space is equal to one. The expanded definition is following: a probability space is a triple $(\Omega,\ \mathcal{F},\ P)$ consisting of:

- the sample space Ω — an arbitrary non-empty set,
- the $\sigma - algebra\ \mathcal{F} \in 2^{\Omega}$ (also called $\sigma - filed$) — a set of subsets of Ω, called events, such that:

  - $\mathcal{F}$ contains the empty set: $\emptyset \in \mathcal{F}$,
  - $\mathcal{F}$ is closed under complements: if $A \in \mathcal{F}$, then also $(\Omega \setminus A) \in \mathcal{F}$,
  - $\mathcal{F}$ is closed under countable unions: if $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$, then also $(\bigcup A_i) \in \mathcal{F}$,

- the probability measure $P : \mathcal{F} \to [0, 1]$ — a function on $\mathcal{F}$ such that:

- P is countably additive: if $\{A_i\} \in \mathcal{F}$ is a countable collection of pairwise disjoint sets, then $P(\bigcup A_i) = \sum P(A_i)$, where $\bigcup$ denotes the disjoint union,

- the measure of entire sample space is equal to one: $P(\Omega) = 1$.

11.146 Probability Axioms

11.146.1 Law of Total Probability

Suppose $B_n (n = 1, 2, 3, \cdots$ is a finite or countably infinity partition of the sample space, then for an event A

$$Pr(A) = \sum_n Pr(A, B_n) \tag{11.114}$$

or,

$$Pr(A) = \sum_n Pr(A|B_n) Pr(B_n) \tag{11.115}$$

It can also be stated for conditional probabilities. Suppose X is an event in the same sample space, we have

$$Pr(A|X) = \sum_n Pr(A|B_n, X) Pr(B_n|X) \tag{11.116}$$

One application of Eq 11.115 is when calculating $Pr(A)$ is difficult, we can introduce an "auxiliary variable" B, in the hope that the conditional probability $Pr(A|B_n)$ is easier to compute.

11.146.2 Law of Total Variance

11.146.3 Law of Total Covariance

11.146.4 Law of Total Expectation

11.146.5 Law of Total Cumulance

11.146.6 Probability Inequalities

11.147 Types of Probabilities: Frequentism and Bayesian

11.148 Random Variables

The Laplace distribution can be written as an infinite mixture of Gaussians with variance w distribution according to an exponential distribution. An exponential distribution can be written as a χ^2 distribution with two degrees of freedom.

11.149 Continuous Random Variables

11.150 Discrete Random Variables

11.151 Joint Distributed Random Variables

11.152 Random Vectors/Matrices

11.153 Function of Random Variables

11.153.1 Transformation

11.153.2 Convolutions: Sum of Normally Distributed Random Variables

11.153.3 Product Distribution

11.153.4 Ratio Distribution

11.154 Useful Distributions

11.155 Discrete Distributions

11.155.1 Poisson Distribution

11.155.2 Bernoulli Distribution

11.155.3 Binomial Distribution

11.155.4 Negative Binomial Distribution

11.155.5 Categorical Distribution

11.155.6 Multinomial Distribution

11.155.7 Geometric Distribution

11.155.8 Hyper-Geometric Distribution

11.155.9 Poisson Distribution

11.156 Continuous Distributions

11.156.1 Uniform Distribution

11.156.2 Exponential Distribution

11.156.3 χ^2 Distribution

11.156.4 Gaussian Distribution

Univariate Gaussian

Multivariate Gaussian

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{11.117}$$

Using the technique of 'completing the square', we have

$$f(\mathbf{x}_1|\mathbf{x}_2) = \frac{f(\mathbf{x}_1, \mathbf{x}_2)}{f(\mathbf{x}_2)} \tag{11.118}$$

$$= \frac{1/\sqrt{2\pi|\mathbf{\Sigma}|}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}}{1/\sqrt{2\pi|\mathbf{\Sigma}_{22}|}\exp\left\{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu})^T\mathbf{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu})\right\}} \tag{11.119}$$

$$= C\cdot\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{11.120}$$

$$= C\cdot\exp\left\{-\frac{1}{2}[(\mathbf{x}_1-\boldsymbol{\mu}_1)^T\mathbf{\Lambda}_{11}(\mathbf{x}_1-\boldsymbol{\mu}_1)+(\mathbf{x}_1-\boldsymbol{\mu}_1)^T\mathbf{\Lambda}_{12}(\mathbf{x}_2-\boldsymbol{\mu}_2)+(\mathbf{x}_2-\boldsymbol{\mu}_2)^T\mathbf{\Lambda}_{21}(\mathbf{x}_1-\boldsymbol{\mu}_1)]\right\} \tag{11.121}$$

$$= C\cdot\exp\left\{-\frac{1}{2}\mathbf{x}_1^T\mathbf{\Lambda}_{11}\mathbf{x}_1+\mathbf{x}_1^T[\mathbf{\Lambda}_{11}\boldsymbol{\mu}_1-\mathbf{\Lambda}_{12}(\mathbf{x}_2-\boldsymbol{\mu}_2)]\right\} \tag{11.122}$$

Hence

$$\mathbf{\Sigma}_{1|2} = \mathbf{\Lambda}_{11}^{-1} \tag{11.123}$$

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 - \mathbf{\Lambda}_{11}^{-1}\mathbf{\Lambda}_{12}(\mathbf{x}_2-\boldsymbol{\mu}_2) \tag{11.124}$$

or

$$\mathbf{\Sigma}_{1|2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} \tag{11.125}$$

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2) \tag{11.126}$$

11.156.5 Dirichlet Distribution

Gamma Function and Beta Function

The *gamma function* is defined for all complex numbers except the non-positive integers. For complex numbers with a positive real part, it is defined via an improper integral that converges:

$$\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}dt \tag{11.127}$$

For all positive numbers z,

$$\Gamma(z) = (z-1)\Gamma(z-1) \tag{11.128}$$

and in particular, if n is a positive integer:

$$\Gamma(n) = (n-1)! \tag{11.129}$$

Note, the gamma function shifts the normal definition of factorial by 1.

The *beta function* is defined by

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt \tag{11.130}$$

for $Re(x), Re(y) > 0$.

It can also be written as

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \tag{11.131}$$

Introduction

The Dirichlet distribution of order $K \geq 2$ with parameters $\alpha_1, \ldots, \alpha_K > 0$ has a probability density function with respect to Lebesgue measure on the Euclidean space $\mathcal{R}^{K-1}$ given by

$$f(x_1, \ldots, x_{K-1}; \alpha_1, \ldots, \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \tag{11.132}$$

for all $x_1, \ldots, x_K > 0$ and $x_1 + \cdots + x_K = 1$. The density is zero outside this open[12] $(K-1)$-dimensional simplex.

The normalizing constant is the multinomial beta function, which can be expressed in terms of the gamma function:

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}, \text{ where } \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \tag{11.133}$$

[12] "Open" here means none of the x_i's can be 1, actually $x_i \in (0,1)$.

11.156.6 t Distribution

11.156.7 Inverse Gaussian Distribution

11.156.8 log-normal Distribution

11.156.9 Laplace Distribution

11.156.10 Beta Distribution

11.156.11 Gamma Distribution

11.156.12 Wishart Distribution

11.157 Quantitative Measure and Characteristic Functions

11.158 Describing Shape of a Distribution: Skewness, Kurtosis

11.159 Describing a Sample: Mean, Variance

11.160 Degrees of Freedom, Mean, Variance, and Moment, Central Moment, Cumulant, Law of the unconscious statistician

11.161 Percentile and Median

11.162 Coefficient of Variation

11.163 Covariance and Correlation

11.164 Moment Generating Function

11.165 Characteristic Function

11.166 Limit Theorems

11.167 Markov and Chebyshev Inequalities

Theorem 11.167.1 Markov's Inequality. If X is a random variable that takes

only nonnegative values, then for any value $a > 0$,

$$P\{X \geq a\} \leq \frac{E[X]}{a}$$

Theorem 11.167.2 Chebyshev's Inequality. If X is a random variable with finite mean μ and variance σ^2, then, for any value $k > 0$,

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

11.168 Weak Law of Large Numbers

Theorem 11.168.1 The Weak Law of Large Numbers. Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables, each having finite mean $E[X_i] = \mu$. Then, for any $\epsilon > 0$,

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right\} \to 0 \text{ as } n \to \infty$$

11.169 Central Limit theorem

Theorem 11.169.1 Central Limit theorem. Let $X_1, X_2, \cdots$ be a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2. Then the distribution of

$$\frac{\frac{X_1 + \cdots + X_n}{n} - \mu}{\sqrt{\sigma^2/n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$.

11.170 Selected Topics of Probability

11.171 Indicator Variables

11.172 Ordered Statistics

11.173 Copula

11.174 coupling

11.175 The Reflection Principle

11.176 Elementary Theory of Stochastic Processes

11.177 Introduction

Definition 11.177.1 A *stochastic process* is a collection of random variables $X(t)$ for $t \in T$, where the *time* parameter T is usually[13] a subset of $\mathcal{R}$. The set $\mathscr{S}$ containing all possible values of $X(t)$ is called the *state space* of the process.

There are four common types of stochastic processes (with corresponding examples):

	Finite or Countable State Space	**Continuous State Space**
Discrete Time	Markov Chains	Harris Chains
Continuous Time	Poisson Processes and Hawkes Processes	Gaussian Processes and Wiener Processes

Table 11.3: Four Common Types of Stochastic Processes.

Definition 11.177.2 A *renewal process* is an idealized stochastic model for events that occur randomly in time (generically called renewals or arrivals). The basic mathematical assumption is that the times between the successive arrivals are independent and identically distributed (i.i.d.).

Definition 11.177.3 A *delayed renewal process* is just like an ordinary renewal process, except that the first arrival time is allowed to have a different distribution than the other inter-arrival times.

Definition 11.177.4 A *counting process* is a stochastic process $\{N(t), t \geq 0\}$ with values that are non-negative, integer, and increasing:

Definition 11.177.5 A *point process* is a collection of mathematical points randomly located on some underlying mathematical space such as the real line, the Cartesian plane, or more abstract spaces.

[13] It can be defined in a general space.

Definition 11.177.6 A *diffusion process* is a solution to a stochastic differential equation. For example, Brownian motion.

Definition 11.177.7 A *jump process* is a type of stochastic process that has discrete movements, called jumps, with random arrival times, rather than continuous movement, typically modeled as a simple or compound Poisson process.

11.178 Markov Chains

11.179 Introduction

Definition 11.179.1 The *Markov property* is defined by the requirement that

$$P(X_{n+1} = x_{n+1}|X_0 = x_0, \cdots, X_n = x_n) = P(X_{n+1} = x_{n+1}|X_n = x_n) \tag{11.134}$$

Definition 11.179.2 The conditional probabilities $P(X_{n+1} = y|X_n = x)$ are called the *transition probabilities* of the chain.

Definition 11.179.3 We say a Markov chain has *stationary transition probabilities* if the transition probabilities $P(X_{n+1} = y|X_n = x)$ is independent of the time n.

From now on, when we say that $X_n, n \geq 0$ forms a Markov chain, we mean that these random variables satisfy the Markov property and have stationary transition probabilities.

Definition 11.179.4 A state a of a Markov chain is called an *absorbing state* if $P(a, a) = 1$ or, equivalently, if $P(a, y) = 0$ for all $y \neq a$.

11.180 Discrete Time Markov Chains

11.180.1 Gambler's Ruin

11.180.2 Discrete Time Branching Processes

11.181 Continuous Time Markov Chains

11.182 Poisson Processes

11.183 Poisson Processes on the Line

11.184 Variable Rate Poisson Processes

11.185 Poisson Processes in Higher Dimensions

11.186 Renewal Theory

11.187 Renewal theory for positive lattice valued random variables as connected with Markov chains: Blackwell's renewal theorem for positive lattice valued random variables

11.188 Selected Topics of Stochastic Processes

11.189 Martingales which are functions of discrete time Markov chains

11.190 Brownian motion: Path properties, reflection principle, random walk approximation.

11.191 Random Fields

11.192 Discrete and Continuous Time Birth and Death processes

11.193 Discrete and Continuous Time Queuing processes

11.194 Finite State Space Pure Jump Processes

11.195 Infinite Server Queue

11.196 Measure-theoretical Probability

11.197 Why Do We Need Rigorous Probability Theory?

called *events*. *Probability measure* is nothing but assigning probability $P(E)$ to each event $E \subset \Omega$, under the following constraints:

1. $P(E) \in [0,1]$
2. $P(\Omega) = 1$
3. $P(\uplus_{i=1}^{\infty} E_i) = \sum_{i=1}^{n} P(E_i)$

where in the last constraint, '$\uplus$' means the union of *disjoint* sets.

One of the motivations[14] of developing measure-theoretical probability theory is triggered by the following question:

Q: *How do we know the three listed axioms are consistent, in particular, the last constraint?*

Actually, the *Banach-Tarski paradox* and the *Vitali paradox* are such counter examples (TBA). We have two choices: either we could reject the *axiom of choice*, which is one of the basic assumptions made in the two paradoxes; or we could claim there are some subsets $E \subset \Omega$ that is 'non-measurable'. In this course, we take the second solution.

11.198 Normal Numbers

Definition 11.198.1 Every $X \in [0,1]$ has a binary decimal expansion $X = 0.d_1d_2d_3\cdots$ or $X = \sum_{k=1}^{\infty} d_k 2^{-k}$, $d_k \in \{0,1\}$. However, it is possible that X may have two expansions, for example, $X = 0.111\cdots = 0.1000\cdots$. We choose the expansion ending in all 1's to make it well-defined.

Definition 11.198.2 $X \in (0,1)$ is normal if $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} d_k = \frac{1}{2}$.

Proposition 11.198.1 If $X \sim U(0,1)$, then $P(X \text{ is normal}) = 1$.

Lemma 11.198.1 Suppose that $X_1, X_2, \cdots$ are i.i.d $Ber(1/2)$. Then $X = \sum_{n=1}^{\infty} X_n 2^{-n}$ is uniform on $(0,1)$.

Proof We (only) need to show $P(X \in (a,b]) = b - a$.

Case 1: $\exists k, s.t. (a,b] = ((k-1)2^{-n}, k2^{n}]$. Since $(k-1)2^{-n} = 0.d_1d_2\cdots d_n000\cdots$,

$$P(X \in ((k-1)2^{-n}, k2^{-n}]) = P(X_1 = d_1, X_2 = d_2, \cdots, X_n = d_n, X_i = 1 for i > n) = 2^{-n} = b - a \tag{11.135}$$

Case 2: $(a,b] = (l2^{-n}, k2^{-n}], l < k$.

$$P(X \in (l2^{-n}, k2^{-n}]) = (k-l)2^{-n} = b - a \tag{11.136}$$

[14] Add unification of discrete and continuous r.v's.

In general, if $a < b$, let $a_n, b_n \in 2^{-n}\mathbf{Z}$ be s.t. $a_n \leq a < a_n + 2^{-n}, b_n - 2^{-n} \leq b < b_n$, then

$$P(X \in (a,b]) \leq P(X \in (a_n, b_n]) = b_n - a_n \tag{11.137}$$

$b_n - a_n - 2^{-n+1} = P(X \in (a_n + 2^{-n}, b_n - 2^{-n}) \leq \epsilon$. Note, $b - a \leq b_n - a_n \leq b - a + 2^{-n+1}$.

We need to show that $P(\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{\infty} = \frac{1}{2}) = 1$. This is the SLLN.

We know that $\{X \in (a,b]\}$ are events. Why is $\{X$ is normal$\}$ an event?

$$\{X \text{ is normal}\} = \left\{ \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{\infty} = \frac{1}{2} \right\} = \bigcap_{l=1}^{\infty} \bigcup_{m=l}^{\infty} \bigcap_{n=m}^{\infty} \left\{ \left| \frac{1}{n}\sum_{k=1}^{\infty} - \frac{1}{2} \right| < \frac{1}{l} \right\} \tag{11.138}$$

This is equivalent to $\forall \epsilon = \frac{1}{l} > 0, \exists m < \infty, s.t., |\frac{1}{n}\sum_{k=1}^{\infty} - \frac{1}{2}| < \epsilon, \forall n \geq m$.

11.199 Formal Definition of Probability Space

Definition 11.199.1 A collection $\mathcal{F}$ of subsets of Ω is a σ-field (or algebra) if

1. $\mathcal{F}$ is non-empty;
2. If $A \subset \mathcal{F}$, then $A^c \in \mathcal{F}$ (closed under complement);
3. If $\{A_i\}$ is a countable sequence of elements of $\mathcal{F}$, then $\cup_i A_i \in \mathcal{F}$ (closed under countable unions).

Note, 1) $\Omega \in \mathcal{F}, \emptyset \in \mathcal{F}$ since $\Omega = A \cup A^c if A \in \mathcal{F}$. $\emptyset = \Omega^c$. 2) $\mathcal{F}$ is closed under countable intersections.

Definition 11.199.2 A *measure space* is a pair $(\Omega, \mathcal{F})$ where $\mathcal{F}$ is a σ-field of subsets of Ω.

Definition 11.199.3 A non-negative measure μ on $(\Omega, \mathcal{F})$ is a function $\mu : \mathcal{F} \to \bar{\mathcal{R}}_+ = [0, \infty]$ s.t.

1. $\mu(\emptyset) = 0$;
2. If A_i is a sequence of disjoint sets in $\mathcal{F}$, then $\mu(\uplus_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

Definition 11.199.4 A *probability space* is a triple $(\Omega, \mathcal{F}, P)$ where $(\Omega, \mathcal{F})$ is a measure space, and P is a measure on $(\Omega, \mathcal{F})$ with $P(\Omega) = 1$.

Example 11.199.1 $\mathcal{F} = 2^{\Omega}$ (all subsets of Ω). If $\Omega = \mathbf{Z}$, then usually, $\mathcal{F} = 2^{\Omega}$ is the σ-field we use. If $\Omega = (0,1]$ or $\mathbf{R}$, then $\mathcal{F}$ is usually too big.

Example 11.199.2 Let $\mathcal{A} \subset 2^{\Omega}$, then $\sigma(\mathcal{A})$ is the smallest σ-field containing $\mathcal{A}$. If $\mathcal{O} \subset \mathbf{R}$ is the collection of all open subsets of $\mathbf{R}$, then $\sigma(\mathcal{O}) = \mathcal{B}$ is called the *Borel σ-field*.

11.200 Lecture 5 (1/21/2015 Wednesday):

(Today and Friday) in Appendix A.

Definition 11.200.1 A non-empty collection of subsets $\mathcal{A} \subset 2^{\Omega}$ is called an algebra if

1. $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$
2. $A, B \in \mathcal{A}$ then $A \cup B \in \mathcal{A}$
3. $A, B \in \mathcal{A}$ then $A \cap B \in \mathcal{A}$

Definition 11.200.2 $\mu : \mathcal{A} \to [0, \infty]$ is a measure on the algebra $\mathcal{A}$ if

1. $\mu(\emptyset) = 0$
2. If $\uplus_{i=1}^{\infty} A_i \in \mathcal{A}$ then $\mu(\uplus_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$

Definition 11.200.3 A measure μ is σ-finite (on an algebra or a σ-field) if $\exists$ a sequence $A_n \nearrow \Omega$ with $\mu(A_n) < \infty$ ($A_n \subset A_{n+1}$ and $\Omega = \cup_{i=1}^{\infty} A_n$)

Theorem 11.200.1 (Caratheodory Extension) Let μ be a σ-finite measure on an algebra $\mathcal{A}$. Then μ has a unique extension to a measure on $(\Omega, \sigma(\mathcal{A}))$.

Note: Meausre on algebras also satisfy Themorem 1.1.1

1. $A \subset B$ then $\mu(A) \leq \mu(B)$
2. $\mu(\cup_{i=1}^{\infty} A_i \leq \sum_i \mu(A_i)$ if $\cup A_i \in \mathcal{A}$

11.201 Probability Space and Measure

11.202 Algebra of Sets

Set Operations

Given two sets $A, B \in \Omega$, there are four basic binary operations on sets:

1. *Union*: $A \cup B = \{x : x \in A \text{ or } x \in B\}$
2. *Intersection*: $A \cap B = \{x : x \in A \text{ and } x \in B\}$
3. *Set difference*: $A \setminus B = \{x : x \in A \text{ and } x \notin B\}$
4. *Set complement*: $A^c = \Omega \setminus A$

Set complement is the "strongest" operation, because if a collection of sets $\mathcal{A}$ is closed under complement, and if it is also closed under any one of the other three operations, $\mathcal{A}$ is closed under the rest two. That is seen from,

$$\text{If closed under } \textit{union} \begin{cases} A \cap B = (A^c \cup B^c)^c \\ A \setminus B = A \cap B^c = (A^c \cup B)^c \end{cases} \tag{11.139}$$

$$\text{If closed under } intersection \begin{cases} A \cup B = (A^c \cap B^c)^c \\ A \setminus B = A \cap B^c \end{cases} \tag{11.140}$$

$$\text{If closed under } difference \begin{cases} A \cap B = A \setminus (A \setminus B) \\ A \cup B = (A^c \cap B^c)^c = (A^c \setminus (A^c \setminus B^c))^c \end{cases} \tag{11.141}$$

The difference operation is the second "strongest" operation, in that if a collection of sets $\mathcal{A}$ is closed under difference, it is closed under intersection, that is seen from,

$$A \cap B = A \setminus (A \setminus B) \tag{11.142}$$

The third "strongest" operation is the union, which can not be implied from difference or intersection, or their combination.

The "weakest" operation is the intersection, which can be implied from the difference.

Class of Set Collection

Definition 11.202.1 Given a set Ω, a non-empty collection $\mathcal{P} \subset 2^\Omega$ is called a *π-system* iff:

$$\forall A, B \in \mathcal{P},\ A \cap B \in \mathcal{P}$$

Notice, this has the weakest requirement.

Definition 11.202.2 Given a set Ω, a non-empty collection $\mathcal{Q} \subset 2^\Omega$ is called a *semiring* iff:

$$\forall A, B \in \mathcal{Q} \text{ and } A \supset B$$

$$\exists C_k \subset \mathcal{Q}, s.t., A \setminus B = \bigcup_{k=1}^{n} C_k$$

Definition 11.202.3 Given a set Ω, a non-empty collection $\mathcal{R} \subset 2^\Omega$ is called a *ring* iff:

$$\forall A, B \in \mathcal{R},\ A \cup B \in \mathcal{R} \text{ and } B \setminus A \in \mathcal{R}$$

There are two points need to mention: the empty set is in a ring, since $A \setminus A = \emptyset$; $\mathcal{A}$ is also closed under intersection (the reverse need not be true).

Definition 11.202.4 A *ring* $\mathcal{A}$ is called an *field* iff $\Omega \in \mathcal{A}$.

So a field is closed under all *finite* combination of set operations.

Definition 11.202.5 An *field* is called a *σ-field* if for any sequence A_n of sets in $\mathcal{A}$, $\cup_{n \geq 1} A_n \in \mathcal{A}$.

∄

11.203 Integration Theory

11.204 Random Variables

11.205 Law of Large Numbers

11.206 Types of Convergence

Convergence in Distribution

Convergence in Probability

Almost Surely Convergence

L^p **Convergence**

11.207 Weak Law of Large Numbers (WLLN)

11.208 Strong Law of Large Numbers (SLLN)

11.209 Central Limit theorem

11.210 Some Tricks

11.211 Prove by Contraposition

11.212 Construct Finer Partition

Given twofinite partitions $\{A_n\}$ and $\{B_m\}$, a finer partition can be constructed as

$$(\cup_{i=1}^{n} A_i) \bigcap (\cup_{j=1}^{m} B_j) = \bigcup (\cap_{i=1}^{n} \cap_{j=1}^{m} A_i B_j)$$

11.213 Prove Equality

To prove two numerical quantities are equal $X = Y$, often times we can do this by showing $X \leq Y$ and $X \geq Y$. Similarly, to prove two sets are equal $E = F$, we can show $E \subset F$ and $E \subset F$.

11.214 An Epsilon of Room

If one has to show that $X \leq Y$, try proving that $X \leq Y + \epsilon, \forall \epsilon > 0$. This trick combines well with the "Prove Equality" trick.

In a similar spirit, if one needs to show that a quantity X vanishes, try showing that $|X| \leq \epsilon, \forall \epsilon > 0$.

If one wants to show that a sequence x_n of real numbers converges to zero, try showing that $\limsup_{n\to\infty} |x_n| \leq \epsilon, \forall \epsilon > 0$

11.215 Interpretations of Probability

11.215.1 Cox's theorem

11.215.2 Principle of Maximum Entropy

11.216 Measure-theoretical Stochastic Processes

11.217 Machine Learning Introduction

There are two main types of Machine Learning problems: the **predictive** or **supervised learning**, and the **descriptive** or **unsupervised learning** [15].

For supervised learning, there are two subtypes: the **classification** problem, where the response variable is categorical (either ordinal or nomial); and the **regression** problem, where the response variable is numerical (either discrete or continuous).

For unsupervised learning, it is also called **density estimation** in Statistics literature. Essentially we want to build models of the form $p(\mathbf{x}_i|\theta)$, and supervised learning can be seen as to build models of the form $p(y_i|\mathbf{x}_i, \theta)$, which is a problem of conditional density estimation. [16]

11.218 Some Distinctions

11.218.1 Machine Learning v.s. Statistical Learning

Machine Learning focuses more on high dimension low noise situations, and performance and efficiency are the main concerns. **Statistical Learning** focuses more on low dimension high noise situations, interpretation and statistical inference are the main concerns.

11.218.2 Parametric v.s. Non-parametric Models

A parametric model has a fixed *finite* number of parameters, while the number of parameters in a non-parametric model grows with the amount of training data. A model with infinite number of parameters is usually non-parametric.

[15] It is also called **Knowledge Discovery** in the Data Mining literature

[16] We see from the formulation that, unsupervised learning usually involves multivariate probability models while supervised learning usually involves univariate probability models.

11.219 A Brief History of Machine Learning

The perceptron model was invented in 1957, and it generated over optimistic view for AI during 1960s. After Marvin Minsky pointed out the limitation of this model in expressing complex functions, researchers stopped pursuing this model for the next decade.

In 1970s, the machine learning field was dormant, when expert systems became the mainstream approach in AI. The revival of machine learning came in mid-1980s, when the decision tree model was invented and distributed as software. It is also in mid 1980s multi-layer neural networks were invented, With enough hidden layers, a neural network can express any function, thus overcoming the limitation of perceptron. We see a revival of the neural network study.

Around 1995, SVM was proposed and have become quickly adopted.

After year 2000, Logistic regression was rediscovered and re-designed for large scale machine learning problems . In the ten years following 2003, logistic regression has attracted a lot of research work.

We discussed the development of 4 major machine learning methods. There are other method developed in parallel, but see declining use today in the machine field: Naive Bayes, Bayesian networks, and Maximum Entropy classifier (most used in natural language processing).

In addition to the individual methods, we have seen the invention of ensemble learning, where several classifiers are used together, and its wide adoption today.

http://en.wikipedia.org/wiki/Timeline_of_artificial_intelligence

1950s-1960s: the Perceptron Model 1970s-1980s: the Expert Systems mid 1980s: Multi-layer Neural Networks 1995: SVM 2000s: Logistic Regression Rediscovered

11.220 Regression v.s. Classification

Consider a supervised learning problem in which we wish to approximate an unknown target function $f : \mathcal{X} \to \mathcal{Y}$, or equivalently $P(Y|X)$. One way to learn $P(Y|X)$ is to use the training data to estimate $P(X|Y)$ and $P(Y)$, and then use Bayes rule to determine $P(Y|X)$.

11.221 Parametric v.s. Nonparametric Models

11.222 Decision Trees

11.223 Clustering

11.224 Dimension Reduction

11.225 Graphical Models

The motivation of graphical models is, in the previous settings, variables are assume to be (conditionally) independent of each other: such as observed variables $\mathbf{x}$ and/or $\mathbf{y}$; or latent variables $\mathbf{z}$ (excluding parameters $\boldsymbol{\theta}$). Graphical models consider the case when variables are correlated.

11.226 Neural Networks

11.227 Kernel Methods

11.228 Support Vector Machines (SVM)

11.229 Gaussian Processes (GP)

11.230 Statistical Learning Theory

Statistical learning theory studies the properties, in particular error-bounds, of learning algorithms in a statistical framework.

The No Free Lunch theorem says, if no assumptions about how training data are related to the testing data, prediction is impossible; furthermore, if no assumptions about the data to be expected, generalization is impossible.

Simply means we need to make the assumption that there is a stationary distribution of data.

Definition 11.230.1 Suppose $f : \mathbb{R} \to \mathbb{R}_+$ and $g : \mathbb{R} \to \mathbb{R}_+$, we write

$$f = O(g) \tag{11.143}$$

if there exists $x_0, \alpha \in \mathbb{R}_+$ such that for all $x > x_0$ we have $f(x) \leq \alpha g(x)$. Replacing $\leq$ with $\geq$, we write $f = \Omega(g)$.

Definition 11.230.2 Suppose $f : \mathbb{R} \to \mathbb{R}_+$ and $g : \mathbb{R} \to \mathbb{R}_+$, we write

$$f = o(g) \tag{11.144}$$

if for every $\alpha > 0$ there exists x_0 such that for all $x > x_0$ we have $f(x) \leq \alpha g(x)$. Replacing $\leq$ with $\geq$, we write $f = \omega(g)$.

Definition 11.230.3 If $f = O(g)$ and $f = \Omega(g)$, then we write $f = \Theta(g)$.

Definition 11.230.4 We write $f = \tilde{O}(g)$ if there exists $k \in \mathbb{N}$ such that $f(x) = O(g(x) \log^k(g(x)))$.

11.231 Latent Dirichlet allocation (LDA)

11.232 Principle Component Analysis (PCA)

11.233 Linear Discriminant Analysis (LDA)

11.234 Expectation Maximization (EM)

11.235 The EM algorithm

11.236 The ECM and ECME algorithms

11.237 The PX-EM algorithm

11.238 Expectation Propagation (EP)

11.239 Markov Chain Monte Carlo Methods

11.240 Introduction

This note is based on Peter Orbanz's BNP notes:

http://people.stat.sc.edu/hansont/stat740/MCMC.pdf

11.241 Notation

Bold upper case letters represent matrices, e.g., $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{\Theta}$. Bold lower case letters represent vector-valued random variables and their realizations (we do not distinguish between the two), e.g., $\mathbf{x}, \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}$. Curly upper case letters represent spaces (i.e., possible values) of random variables, e.g., $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \Theta$.

11.242 Introduction

Markov chain Monte Carlo (MCMC) methods can be used to draw random samples from a target distribution p. It is particularly useful in Bayesian data

analysis, due to the difficulties of evaluating the denominator in the Bayes' formula, a.k.a. the partition function.

1. A discrete-time, discrete-space Markov chain is $X^{(0)}, X^{(1)}, \ldots$ where $X^{(t)}$ obeys the Markov property that

$$P\left[X^{(t)}\middle|x^{(0)},\ldots,x^{(t-1)}\right] = P\left[X^{(t)}\middle|x^{(t-1)}\right] \tag{11.145}$$

2. A Markov chain is *irreducible* if any state j can be reached from any state i in a finite number of steps for all i and j.

3. A Markov chain is *periodic* if it can visit certain portions of the state space only at regularly spaced intervals.

The MCMC sampling strategy is to construct an irreducible, aperiodic Markov chain for which the stationary distribution equals the target distribution p.

Suppose we want to draw samples from $p(x)$. The M-H algorithm proceeds as follows: Draw a candidate state, x^*, according to the proposal distribution $g(x^*|x)$, by computing the acceptance probability

$$\alpha(x^*, x) = \min\left[1, a(x^*, x) = \frac{p(x^*)g(x|x^*)}{p(x)g(x^*|x)}\right]. \tag{11.146}$$

where $a(x^*, x)$ is called the M-H ratio, and $\alpha(x^*, x)$ the probability of move. With *probability of move* $\alpha(x^*, x)$, set the new state, x' to x^*. Otherwise, let x' be the same as x. The intuition behind the probability of move is that, if the detailed balance condition is satisfied: $p(x)g(x^*|x) = p(x^*)g(x|x^*)$, then we are done, otherwise, the denominator $g(x^*|x)p(x)$ is proportional to the probability of moving from x to x^*, if it is large then the numerator, which is proportional to the probability of moving from x^* to x, then we should penalize it.

The sampled sequence may contain duplicated copies of data points, the frequency of which is used to correct the difference between the proposal distribution and the target one. A well chosen proposal distribution produces candidate values that efficiently cover the support of the target distribution.

11.243 Independence Chains

If we choose the proposal distribution to be

$$g(x^*|x) = g(x^*) \tag{11.147}$$

then the M-H ratio is

$$a(x^*, x) = \frac{p(x^*)g(x)}{p(x)g(x^*)}. \tag{11.148}$$

For example, in a Bayesian framework, if the target distribution is the posterior $p(\theta|\mathbf{y})$, where $\mathbf{y}$ is the data. Then, if we choose the proposal distribution to be the prior $p(\theta)$

$$g(\theta^*|\theta) = p(\theta^*) \tag{11.149}$$

then the M-H ratio is

$$a(\theta^*, \theta|\mathbf{y}) = \frac{p(\theta^*|\mathbf{y})p(\theta)}{p(\theta|\mathbf{y})p(\theta^*)} = \frac{p(\mathbf{y}|\theta^*)p(\theta^*)/p(\mathbf{y})}{p(\mathbf{y}|\theta)p(\theta)/p(\mathbf{y})} \frac{p(\theta)}{p(\theta^*)} = \frac{p(\mathbf{y}|\theta^*)}{p(\mathbf{y}|\theta)} \tag{11.150}$$

So if the proposal distribution is the prior, the M-H ratio is the likelihood ratio.

11.244 Random walk chains

Let x^* be generated by setting

$$x^* = x + \epsilon, \quad \epsilon \sim h(\epsilon) \tag{11.151}$$

or equivalently,

$$g(x^*|x) = h(x^* - x) \tag{11.152}$$

For example, h can be the uniform, or the standard normal, or the Student's t distribution.

11.245 Gibbs sampler

Suppose it is easy to sample from the univariate conditional distributions:

$$x_i|\mathbf{x}_{-i} \sim f(x_i|\mathbf{x}_{-i}) \tag{11.153}$$

then the basic Gibbs sampler can be described as follows:

1. Select starting values $x^{(0)}$ and set $t = 0$.
2. Generate, in turn for $i = 1, \ldots, n$:

$$x_i{}^{(t+1)}|\mathbf{x}_{-i}^{(t)} \sim f\left(x_1|\mathbf{x}_{-i}^{(t)}\right). \tag{11.154}$$

3. Increment t and go to step 2.

A hybrid MCMC may contain different types of samplers. For example, The M-H within Gibbs algorithm is typically useful when the univariate conditional density for one or more elements is not available in closed form.

11.246 Test for Convergence

1. Burn-in.
2. Run multiple chains, and if the within- and between-chain behaviors are similar, suggests that the chains are stationary. Gelman-Rubin statistic.
3. Plot samples against time, or log-likelihood against time.
4. Autocorrelation function (ACF) plot: lag versus correlation. Slow decay suggests poor mixing.
5. Re-parameterize the model may help.
6. Burn-in should be about 5000 iterations, chain lengths should be about 100 times the burn-in.
7. Standard error should be less than 5% of the standard deviation.

11.247 How it is used

Marginalization: just ignore others. Mean and variance: use samples. Probability estimates: estimated by the frequencies. Standard error of estimates: batch runs to obtain estimates and compute mean and standard error (divided by the square root of batch size). Density: kernel density or simply histogram.

11.248 Advanced MCMC methods

Slice sampling and other auxiliary variable methods, reversible jump MCMC, perfect sampling, Hit-and-run (choose a direction and then a distance to run), multi-try (choose from a set of candidates), Langevin M-H (random walk with drift) and etc.

11.248.1 Slice sampling

Introduce an auxilary varialbe u, and if we can sample from $f(x,u) = f(x)f(u|x)$ then dropping u and retain x as desired. The slice sampling works as follows:

$$u^{(t+1)}|x^{(t)} \sim \text{Unif}\left(0, f\left(x^{(t)}\right)\right) \tag{11.155}$$

$$x^{(t+1)}|u^{(t+1)} \sim \text{Unif}\left(x : f(x) \geq u^{(t+1)}\right) \tag{11.156}$$

It is particularly useful for multi-modal problems (but not for high dimensional ones).

11.248.2 Reversible Jump MCMC

RJMCMC is suitable for nonparametric models where model dimensions change. The key is to use auxiliary variables to match the dimensions.

11.249 Introduction

11.249.1 Sampling Methods in General

Inverse Transform Sampling

Importance Sampling

11.249.2 Rejection Sampling

Suppose we want to sample from $P(X)$ by utilizing a *proposal distribution* $Q(X)$, from which we can take samples easier. Let $C = \sup\left\{\frac{P(x)}{Q(x)}, \forall x\right\}$, so $\frac{P(x)}{CQ(x)} \leq 1, \forall x$.

We propose a new value x' from $Q(X)$ and accept this new value with probability

$$A(x'|x) = \frac{P(x')}{CQ(x')} \tag{11.157}$$

This is called the *rejection sampling* method.

Remark 11.249.1 If we plug in Equation 11.157 into Equation 11.169, then

$$LHS = P(x)Q(x'|x)\frac{P(x')}{CQ(x')} = P(x)Q(x')\frac{P(x')}{CQ(x')} = \frac{1}{C}P(x)P(x') \tag{11.158}$$

$$RHS = P(x')Q(x|x')\frac{P(x)}{CQ(x)} = P(x')Q(x)\frac{P(x)}{CQ(x)} = \frac{1}{C}P(x')P(x) \tag{11.159}$$

which shows that the rejection sampling ratio in Equation 11.157 is a special solution to the detailed balance equation.

11.250 Markov Chains

Definition 11.250.1 A *Markov chain* is a collection of random variables $\{X^{(0)}, X^{(1)}, X^{(2)}, \cdots\}$, satisfying the Markov property

$$P(X^{(n+1)}|X^{(n)}, \cdots, X^{(0)}) = P(X^{(n+1)}|X^{(n)}) \tag{11.160}$$

Definition 11.250.2 A distribution over the states of a homogeneous[17] Markov chain is *invariant* (or *stationary*) with respect to transition probabilities T if

$$\pi = \pi T \tag{11.161}$$

[17] The transition matrix is invariant of time.

A Markov chain can have more than one invariant distribution. If T is the identity matrix, for example, then any distribution is invariant.

We are interested in designing a Markov chain (its initial probabilities and transition matrix) for which the distribution we wish to sample from, given by π, is invariant. Hence, if we run the chain for sufficient time, it will converge to π, and then we can sample from it and compute the quantities desired.

Why do we need detailed balance?

Definition 11.250.3 Often, we will use *time reversible* homogeneous Markov chains that satisfy the more restrictive condition of *detailed balance*:

$$\pi(x)T(x,x') = \pi(x')T(x',x) \tag{11.162}$$

Remark 11.250.1 One might be tempted to conclude that the detailed balance condition always holds, since

$$\pi(x)T(x,x') = P(x)P(x'|x) = P(x',x) \tag{11.163}$$
$$\pi(x')T(x',x) = P(x')P(x|x') = P(x,x') \tag{11.164}$$

and $P(x',x) = P(x,x')$.

The problem here is, x and x' denote different values, not different random variables. That is, in the argument of $\pi(\cdot)$ or in the first argument of $T(\cdot,\cdot)$, it is the value of the random variable X_0 (stochastic process at current time); and in the second argument of $T(\cdot,\cdot)$, it is the value of the random variable X_1 (stochastic process at the next time step). For example, we may have

$$\pi(X_0=1)T(X_0=1,X_1=2) = P(X_0=1)P(X_1=2|X_0=1) = P(X_1=2,X_0=1) \tag{11.165}$$

and

$$\pi(X_0=2)T(X_0=2,X_1=1) = P(X_0=2)P(X_1=1|X_0=2) = P(X_1=1,X_0=2) \tag{11.166}$$

which are usually not the same. Here $x=1$ and $x'=2$.

Note, the detailed balance condition implies π is an invariant distribution:

$$\sum_{i=1}^{n} \pi(x_i)T(x_i,x) = \sum_{i=1}^{n} \pi(x)T(x,x_i) = \pi(x)\sum_{i=1}^{n} T(x,x_i) = \pi(x) \tag{11.167}$$

It is possible for a distribution to be invariant without detailed balance holding. For example, the uniform distribution ($= 1/3$) on the state space $\{0,1,2\}$ is invariant with respect to the homogeneous Markov chain with transition probabilities $T(0,1) = T(1,2) = T(2,0) = 1$ and all others zero, but detailed balance does not hold.

Definition 11.250.4 A Markov chain is *ergodic* (or *irreducible*) if it is possible to go from every state to every state (not necessarily in one move).

Theorem 11.250.1 Let T be the transition matrix for a regular chain. Then as $n \to \infty$, the powers T^n approach a limiting matrix W with all rows the same vector $\mathbf{w}$. The vector $\mathbf{w}$ is a strictly positive probability vector (i.e., the components are all positive and they sum to one).

11.251 Metropolis-Hastings Algorithm

Suppose we can evaluate a distribution $P(X)$, but lack of information about how to directly draw samples from it. We wish to construct a Markov chain, utilizing the detailed balance condition (Equation 11.162), such that the induced invariant distribution is nothing but $P(X)$.

Assume we can take a new sample x' from $Q(x'|x)$, a known *proposal distribution* conditioned on the current state x of the Markov chain, and then we accept this new value x' based on a to-be-determined probability $A(x'|x)$. We define a Markov chain based on this procedure,

$$T(x, x') = Q(x'|x)A(x'|x) \tag{11.168}$$

If this Markov chain further satisfies the detailed balance condition (Equation 11.162),

$$P(x)Q(x'|x)A(x'|x) = P(x')Q(x|x')A(x|x') \tag{11.169}$$

then we can take

$$A(x'|x) = \min\left\{\frac{P(x')Q(x|x')}{P(x)Q(x'|x)}, 1\right\} \tag{11.170}$$

The ratio in the "min" function is known as the "Hastings ratio" (or acceptance ratio).

Now, running the chain for some sufficient time, we would obtain samples from the desired distribution $P(X)$, which is designed to be the same as the induced invariant distribution.

Remark 11.251.1 The $A(x'|x)$ defined above satisfies the detailed balance condition, since

$$P(x)Q(x'|x)A(x'|x) = P(x)Q(x'|x)\min\left\{\frac{P(x')Q(x|x')}{P(x)Q(x'|x)}, 1\right\} = \min\left\{P(x')Q(x|x'), P(x)Q(x'|x)\right\} \tag{11.171}$$

$$P(x')Q(x|x')A(x|x') = P(x')Q(x|x')\min\left\{\frac{P(x)Q(x'|x)}{P(x')Q(x|x')}, 1\right\} = \min\left\{P(x)Q(x'|x), P(x')Q(x|x')\right\} \tag{11.172}$$

and $\min\{P(x')Q(x|x'), P(x)Q(x'|x)\} = \min\{P(x)Q(x'|x), P(x')Q(x|x')\}$.

11.251.1 Metropolis Algorithm

In the Hastings ratio, if the proposal distribution is symmetric $Q(x|x') = Q(x'|x)$, such as a Gaussian distribution, then Equation 11.170 becomes

$$A(x'|x) = \min\left\{\frac{P(x')Q(x|x')}{P(x)Q(x'|x)}, 1\right\} = \min\left\{\frac{P(x')}{P(x)}, 1\right\} \tag{11.173}$$

this special case is called the "Metropolis Algorithm".

11.251.2 Gibbs Sampling

Suppose we wish to sample a random vector $\mathbf{X} = (X_1, \cdots, X_n)$, and its full conditional distribution $Q(X_i|\mathbf{X}_{-i})$ is known, where $\mathbf{X}_{-i} = (X_1, \cdots, X_{i-1}, X_{i+1}, X_n)$. Then if we sample X_i component-wise from $Q(\mathbf{x}'|\mathbf{x}) = Q(x_i'|\mathbf{x}_{-i})$, and bearing in mind that $\mathbf{x}'_{-i} = \mathbf{x}_{-i}$ when sample x_i, the Hastings ratio becomes,

$$\begin{aligned}
\frac{P(\mathbf{x}')Q(\mathbf{x}|\mathbf{x}')}{P(\mathbf{x})Q(\mathbf{x}'|\mathbf{x})} &= \frac{P(x_i'|\mathbf{x}'_{-i})P(\mathbf{x}'_{-i})Q(x_i|\mathbf{x}'_{-i})}{P(x_i|\mathbf{x}_{-i})P(\mathbf{x}_{-i})Q(x_i'|\mathbf{x}_{-i})} && (11.174)\\
&= \frac{P(x_i'|\mathbf{x}_{-i})P(\mathbf{x}_{-i})Q(x_i|\mathbf{x}_{-i})}{P(x_i|\mathbf{x}_{-i})P(\mathbf{x}_{-i})Q(x_i'|\mathbf{x}_{-i})} && (11.175)\\
&= 1 && (11.176)
\end{aligned}$$

11.251.3 Collapsed Gibbs Sampling

11.251.4 Metropolis-Within-Gibbs

If not all full conditional probabilities are known or can be easily sampled from, we can sample those random variables with known conditional probabilities using the Gibbs algorithm, and others using Metropolis-Hastings algorithm. This is called *Metropolis-Within-Gibbs*.

11.252 Slice Sampling

11.252.1 Elliptical Slice Sampling

11.253 Split-Merge Sampling

11.254 Hamiltonian Monte Carlo

11.255 Data Fusion and Particle Filter (Sequential MCMC)

11.256 Reversible jump MCMC

11.257 Convergence Diagnostics

11.258 Bayesian Nonparametrics

The concept of functions can be generalized to that of algorithms, which describe procedures, with loops and conditional tests, of how to generate output from input.

A draw from a finite dimensional Gaussian distribution is a real number, while a real-valued function can be considered as a sequence of (uncountably) infinite number of real numbers. A Gaussian Process (GP) is an infinite dimensional generalization of a Gaussian distribution. It defines a prior over real-valued functions, and a sample of it is a particular example of such functions.

A draw from a finite dimensional Dirichlet distribution is a (discrete) probability measure. A Dirichlet Process (DP) is an infinite dimensional generalization of a Dirichlet distribution. It defines a prior over probability measures, and a sample of it is a probability measure. Distributions drawn from a Dirichlet process are discrete, but cannot be described using a finite number of parameters, thus the classification as a nonparametric model.

Note, that we do not have a measurement of the function, as in the GP case but a sample of the true probability measure; this is the main difference between GP and DP.

11.259 Introduction

This note is based on Peter Orbanz's BNP notes:

http://stat.columbia.edu/~porbanz/npb-tutorial.html

11.260 Notation

Bold upper case letters represent matrices, e.g., $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{\Theta}$. Bold lower case letters represent vector-valued random variables and their realizations (we do not distinguish between the two), e.g., $\mathbf{x}, \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}$. Curly upper case letters represent spaces (i.e., possible values) of random variables, e.g., $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \Theta$.

11.261 Terminology

11.261.1 Parametric and nonparametric models

In a set of probability spaces $\{(\mathcal{Y}, \mathcal{F}, \mathcal{P}_\Theta)\}$, a *statistical model* $\mathcal{M}$ on a sample space $\mathcal{Y}$ is a set of probability measures $\mathcal{P}_\Theta$ on $\mathcal{Y}$. If we write $PM(\mathcal{Y})$ for the space of all probability measure on $\mathcal{Y}$, a model is a subset $\mathcal{M} \subset PM(\mathcal{Y})$. Every element of $\mathcal{M}$ has a one-to-one mapping (hence the model is *identifiable*) with its parameter $\boldsymbol{\theta}$ with values in a parameter space Θ, that is,

$$\mathcal{M}(\mathbf{y}) = \{P_{\boldsymbol{\theta}}(\mathbf{y}) | \boldsymbol{\theta} \in \Theta\}, \quad \mathbf{y} \in \mathcal{Y}. \tag{11.177}$$

For example, a first order polynomial is a model, and a second order polynomial is another model. We can of course fit a model to the observed data, but *model* itself is an abstract concept, where the parameter values of a model need not be specified. We call a model *parametric* if Θ has finite dimension, and *nonparametric* if Θ has infinite dimension.

To formulate statistical problems, we assume that n observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$ with values in $\mathcal{Y}$ are observed, which are drawn i.i.d. from a measure $P_{\boldsymbol{\theta}}$ in the model, i.e.,

$$\mathbf{y}_1, \ldots, \mathbf{y}_n \sim_{iid} P_{\boldsymbol{\theta}} \qquad \text{for some } \boldsymbol{\theta} \in \Theta \tag{11.178}$$

The objective of statistical *inference* is then to draw conclusions about the value of $\boldsymbol{\theta}$ (and hence about the distribution $P_{\boldsymbol{\theta}}$ of the data) from the observations.

11.261.2 Bayesian and Bayesian nonparametric models

In Bayesian statistics, all parameters are considered as random variables. Hence under a Bayesian model, data are generated in two stages, i.e.,

$$\boldsymbol{\theta} \sim P(\boldsymbol{\theta}) \tag{11.179}$$

$$\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \boldsymbol{\theta} \sim_{iid} P_{\boldsymbol{\theta}}(\mathbf{y}) \tag{11.180}$$

The objective is then to determine the *posterior distribution* – the conditional distribution of $\boldsymbol{\theta}$ given the observed data,

$$\pi(\boldsymbol{\theta} | \mathbf{y}_1, \ldots, \mathbf{y}_n) \tag{11.181}$$

A *Bayesian nonparametric* model is a Bayesian model whose parameter space Θ has infinite dimension. To define a Bayesian nonparametric model, we have to define a prior π on an infinite-dimensional space, which is a stochastic process with paths (i.e. realizations) in Θ.

11.262 Clustering and the Dirichlet process

11.262.1 Finite mixture models

The basic assumption of a clustering problem is that each observation $\mathbf{y}_i$ belongs to a single cluster $k \in \{1, \cdots, K\}$, which has a cluster distribution

$$P_k(\mathbf{y}_i | z_i = k) \tag{11.182}$$

where we have defined a latent variable z_i, indicating the cluster assignment of observation $\mathbf{y}_i$. Note that under the Bayesian framework, the latent variable z_i itself has a distribution

$$p_k^i \equiv P(z_i = k) \tag{11.183}$$

The marginal distribution of the observation $\mathbf{y}_i$ is then

$$P(\mathbf{y}_i) = \sum_{k=1}^{K} P(z_i = k) P_k(\mathbf{y}_i | z_i = k) \tag{11.184}$$

A model of this form is called a *finite mixture model*.

11.262.2 Bayesian mixture models

Suppose we know there are K clusters, we first sample the cluster parameters from some base measure:

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K \sim_{iid} G(\boldsymbol{\beta}) \tag{11.185}$$

We then independently sample the latent cluster assignment vectors and the actual observations:

$$(p_1^i, \ldots, p_K^i) \sim \text{Dirichlet}_K(\alpha) \tag{11.186}$$

$$z_i \sim \text{Categorical}(p_1^i, \ldots, p_K^i) \tag{11.187}$$

$$\mathbf{y}_i \sim P_k(\mathbf{y}_i | \boldsymbol{\theta}_k, z_i = k) \tag{11.188}$$

11.262.3 Dirichlet Process

Definition 11.262.1 If $\alpha > 0$ and if G is a probability measure on Ω_ϕ, the random discrete probability measure Θ generated by

$$V_1, V_2, \cdots \sim_{iid} \text{Beta}(1, \alpha) \tag{11.189}$$

$$C_k = V_k \prod_{j=1}^{k-1} (1 - V_k) \tag{11.190}$$

$$\Phi_1, \Phi_2, \cdots \sim_{iid} G \tag{11.191}$$

is called a *Dirichlet process (DP)* with base measure G and concentration α, and denote its law by $\text{DP}(\alpha, G)$.

11.263 Glossary

http://alumni.media.mit.edu/~tpminka/statlearn/glossary/

- Model Evidence: Model evidence, or *marginal likelihood*, or *normalization constant*, is a likelihood function in which all parameter variables have been marginalized, usually denoted by $P(D|M)$, or $P(D)$ for short. For example, in polynomial regression, M may denotes the degree of the regression function, and parameter variables are the coefficients for any given M.

- Filtering is the task of tracking the posterior distribution of the latent state variable in state space models.

11.264 Useful Resources

11.265 Data Sets

11.266 Packages and Source Codes

11.267 Important Papers

Part IV

Numerical Methods and Optimization

Chapter 12

Optimization Introduction

Quadratic optimization problems (including, e.g., least-squares) form the base of the hierarchy; they can be solved exactly by solving a set of linear equations. *Newton's method* is the next level in the hierarchy. In Newton's method, solving an unconstrained or equality constrained problem is reduced to solving a sequence of quadratic problems. *The interior-point methods*, which form the top level of the hierarchy, solve an inequality constrained problem by solving a sequence of unconstrained, or equality constrained, problems. Besides Newton's method, there are quasi-Newton, conjugate-gradient, bundle, cutting-plane algorithms, and etc.

Optimization problems can be broadly divided into two types: linear optimization and nonlinear optimization, the later of which consists of unconstrained and constrained optimization problems. Obtaining necessary and sufficient conditions is one of the central problems of nonlinear optimization. The main theory is the study of Lagrange multipliers, including the Karush-Kuhn-Tucker (KKT) theorem and its extensions. However, the theory of Lagrange multipliers is far from adequate, since it does not take into account the difficulties associated with solving the equations resulting from the necessary conditions.

Definition 12.0.1 *Global convergence analysis.* The verification that a given algorithm will in fact generate a sequence that converges to a solution point. *Local convergence analysis* or *complexity analysis.* The rate at which the generated sequence of points converges to the solution.

Chapter 13

Linear Programming

13.1 Basic Properties of Linear Programs

Definition 13.1.1 Let A be an $m \times n$ matrix and B be any *nonsingular* $m \times m$ sub-matrix made up of columns of A. Then, if all $n - m$ components of x not associated with columns of B are set equal to zero, the solution to the resulting set of equations is said to be a *basic solution* with respect to the basis B. The components of x associated with columns of B are called *basic variables*.

Definition 13.1.2 If one or more of the basic variables in a basic solution has value zero, that solution is said to be a *degenerate basic solution*.

Definition 13.1.3 A feasible solution that is also basic is said to be a *basic feasible solution*; if this solution is also a degenerate basic solution, it is called a *degenerate basic feasible solution*.

Theorem 13.1.1 Fundamental Theorem of Linear Programming. Given a linear program in standard form. i) if there is a feasible solution, there is a basic feasible solution; ii) if there is an optimal feasible solution, there is an optimal basic feasible solution.

Simplex algorithm has an exponential worst-case complexity, but polynomial average-case complexity.

Chapter 14

Unconstrained Optimization

14.1 Univariate Problems (Bisection, Newton, Secant Methods)

Note that, $\min(f(x)))$ can be converted to the root-finding $f'(x) = 0$ problem.

14.1.1 Bisection Method

Suppose g' is continuous on $[a_0, b_0]$ and $g'(a_0)g'(b_0) \leq 0$, then the Intermediate Value Theorem implies that there exists at least one x^* for which $g'(x^*) = 0$ and hence x^* is a local optimum of g. To find this local optimum, the Bisection Method systematically halves the interval at each iteration, by checking the product of g'.

The updating equations are

$$[a_{t+1}, b_{t+1}] = \begin{cases} [a_t, x^{(t)}], & \text{if } g'(a_t)g'(x^t) \leq 0 \\ [x^{(t)}, b_t], & \text{if } g'(a_t)g'(x^t) > 0 \end{cases} \tag{14.1}$$

and $x^{t+1} = \frac{a_{t+1}+b_{t+1}}{2}$.

14.1.2 Newton's Method

Suppose g twice differentiable. At iteration t, Newton's method approximates $g'(x^*)$ by the linear Taylor series expansion:

$$g'(x^*) = g'(x^{(t)}) + (x^* - x^{(t)})(g''(x^{(t)})) \tag{14.2}$$

which gives us

$$x^* = x - \frac{g'(x^{(t)})}{g''(x^{(t)})} \tag{14.3}$$

14.1.3 The Secant Method

Recall that the Newton's method requires the function's derivative, which is always available. The secant method approximates the derivative with difference. It works as follows:

- Start with two approximations x_0 and x_1.
- Compute the $(k+1)$th approximation with

$$x_{k+1} \equiv x_k - f(x_k)/\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} \tag{14.4}$$

- The convergence rate of the secant method is 1.618.

14.2 Quasi-Newton Methods

Chapter 15

Convex Optimization

Part V

Software Engineering and Algorithms

Part VI

Interview Questions

Chapter 16

LeetCode

Chapter 17

Kaggle

Chapter 18

FinancialMathematicsProblems

Chapter 19

FinancialStatisticsProblems

Chapter 20

FinancialProgrammingProblems

Chapter 21

BrainTeasers

21.1 Q & A

Question 21.1.1 Why is the difference between Statistics and Machine Learning?

Answer 21.1.1 Statistics focuses on interpretability while Machine Learning cares more about predictability.

Question 21.1.2 Why is the name "statistical" machine learning / data mining / pattern recognition?

Answer 21.1.2 It means the data is in vector form and not in, for example, strings, where it will be called "syntactical/structural" pattern recognition.

Question 21.1.3 How can I categorize machine learning models and methods?

Answer 21.1.3 In general, machine learning models can be categorized according to their strategies for generating features: – fixed basis function models - basis functions are pre-designed; – adaptive basis function models (CART, Neural Networks) - form or parameters of basis functions learned from data; – kernel models (SVM, GP) - basis functions are implicitly defined, dimension of which is essentially infinite; – latent variable models / dimension reduction (HMMs, Sate Space Models; PCA, ICA).

Question 21.1.4 How can I sample a random variable x marginally if I know how to sample jointly $p(x, y)$?

Answer 21.1.4 Simply discard y and keep x in the joint samples.

Question 21.1.5 What's the difference between /learning/ and /inference/.

Answer 21.1.5 Learning is to fit parameters θ to a set of observations (x_i, y_i) while inference is to identify the input x for a particular observation y, using the learned parameters θ. EM can be thought of as iterating between learning and inference.

Question 21.1.6 What is curse of dimensionality?

Answer 21.1.6 The curse of dimensionality is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality. Also, organizing and searching data often relies on detecting areas where objects form groups with similar properties; in high dimensional data, however, all objects appear to be sparse and dissimilar in many ways, which prevents common data organization strategies from being efficient.

Question 21.1.7 Why over-fitted polynomial models (without regularization) tend to have larger coefficients?

Answer 21.1.7 Because coefficients essentially are measures of "derivatives" of different orders, the larger the coefficients, the larger the derivatives, hence the more the function changes.

Question 21.1.8 What is bias-variance trade-off, and its solution?

Answer 21.1.8

$$E[(y-\hat{f})^2] = [Bias(\hat{f})]^2 + Var(\hat{f}) + \sigma^2 \tag{21.1}$$

where $Bias(\hat{f}) = E[\hat{f} - f]$ and $Var(\hat{f}) = E[(\hat{f} - E(\hat{f}))^2] = E[\hat{f}^2] - (E[\hat{f}])^2$.

One way of resolving the trade-off is to use mixture models and ensemble learning. For example, boosting combines many "weak" (high bias) models in an ensemble that has lower bias than the individual models, while bagging combines "strong" learners in a way that reduces their variance.

Question 21.1.9 What are parametric models, non-parametric models, and etc.?

Answer 21.1.9 A **parametric model** $\mathcal{M}$ is a collection of probability distributions $P_{\boldsymbol{\theta}}$, each of which is described by a *finite dimensional* (vector) parameter $\boldsymbol{\theta}$. A parametric model is called identifiable if the mapping $\boldsymbol{\theta} \to P_{\boldsymbol{\theta}}$ is invertible.

$$\mathcal{M} = \{P_{\boldsymbol{\theta}} | \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k\} \tag{21.2}$$

A **non-parametric model** may refer to two interpretations: 1) it may refer to models that do not rely on data belonging to any particular distribution[1], but reply on comparative properties (statistics) of the data, or population, such as the "order statistics", or 2) it may refer to models that do not assume the structure of a model is fixed, i.e., the model grows in size to accommodate the complexity of the data. In these techniques, individual variables are typically assumed to belong to parametric distributions, and assumptions about the types of connections among variables are also made.

[1] Distribution-free methods are such examples, but they are not equivalent concepts.

Question 21.1.10 What are generative models, discriminative models?

Answer 21.1.10 For a supervised learning problem in which we wish to approximate an unknown target function $f : \mathcal{X} \to \mathcal{Y}$, or equivalently $P(Y|X)$, one approach is to model $P(Y|X)$ directly, which is called a discriminative model; another approach is to model $P(X, Y)$, or equivalently $P(Y)$ and $P(X|Y)$, and then use the Bayes' rule, to obtain $P(Y|X)$ for each $X = x$ query.

Question 21.1.11 What are log-linear models?

Answer 21.1.11 Such models have many names, including maximum-entropy models, exponential models, and Gibbs models; Markov random fields are structured log-linear models, conditional random fields are Markov Random Fields with a specific training criterion.

Question 21.1.12 Bayesian v.s. Frequentism?

Answer 21.1.12 There are two main schools of statistical inference: Frequentist and Bayesian[2]. The controversies arise when it comes to how to interpret the randomness of data point generating process.

Bayesians consider parameters to be *random*, and the observed data are conditioned on a realization of such random variables. Notice the natural hierarchical structure in this interpretation. The goal is to inference $p(\boldsymbol{\theta}|D_{\boldsymbol{\theta}_0})$, where $\boldsymbol{\theta}$ are the parameters and $D_{\boldsymbol{\theta}_0}$ the observed data set conditioned on realized values $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$. Frequentists consider parameters to be *unknown but fixed*, and the observed data set is just a sample from the population. As in [67], Efron said "... Bayesian averages involve only the data value $\bar{x}$ actually seen, rather than a collection of theoretically possible other $\bar{x}$ values."

Also, as answered by Michael Hochster in [63] and [71]: Suppose h is the unknown constant, and H is the statistic computed from a sample. For Frequentists, it is valid to write $P(L \leq h \leq U) = 95\%$ or $P(70 \leq H \leq 74) = 95\%$, but not $P(70 \leq h \leq 74) = 95\%$ (this is 0 or 1). So the correct way to say is either "if the same experiment procedure is repeated 100 times, 95 times of the CIs will cover the unknown true value h", or "before the experiment, the probability is 95% that the CI to be obtained will cover h".

Wasserman said in [62] that the two schools of inference differ in their *goals*, not the *methods*: the goal of Frequentist inference is to construct procedures with frequency guarantees, and the goal of Bayesian inference is to quantify and manipulate degrees of beliefs.

Further Readings: Stein's example, Likelihood principle [64], [65], [69], [70].

Question 21.1.13 What are Learning, Machine Learning, Data Mining, Statistics, and their differences?

Answer 21.1.13 Learning is a process of improving performance with experience. There are two main types of learning: deductive learning and inductive learning. Deductive learning learns to apply generalization concepts (rules) to

[2]Another being Fiducial inference, or Fisherian inference.

examples; inductive learning learns to generalized concepts (rules) from examples.

Machine Learning is an example of inductive learning of machines (computers).

A big difference between Machine Learning and Data Mining is Reinforcement Learning. While one of the main goals of statistics is hypothesis testing, one of the main goals of data mining is the construction of hypotheses.

Question 21.1.14 What is the difference between learning and inference?

Answer 21.1.14 Inference reasons about unknown probability distributions; (parameter) learning is finding point estimates of quantities in the model. In Statistics, no distinction between learning and inference only inference (or estimation); and in Bayesian Statistics, all quantities are probability distributions, so there is only the problem of inference. Inference in the Machine Learning community also includes making predictions.

So your inference algorithm gives you posteriors in functional forms, and learning algorithm estimates parameter values from data, and you then inference about predictions using the fitted model.

Part VII

Notes

Bibliography

[1] James Stewart *Calsulus - Early Transcendentals*. Cengage Learning, 2012

[2] Walter Rudin *Principles of Mathematical Analysis*. McGraw-Hill Companies, Inc., 1976.

[3] H. L. Royden *Real Analysis*. Pearson Eduction, Inc., 1988.

[4] Erwin Kreyszig *Introductory Functional Analysis with Applications*. Wiley, 1989.

[5] Gerald B. Folland *Real Analysis: Modern Techniques and Their Applications*. Wiley, 1999.

[6] Alberto Torchinsky *Real Variables*. Westview Press, 1995.

[7] http://normaldeviate.wordpress.com/2012/11/17/what-is-bayesianfrequentist-inference/

[8] http://www.quora.com/What-is-the-difference-between-Bayesian-and-frequentist-statisticians

[9] http://www.bayesian-inference.com/advantagesbayesian

[10] http://www.bayesian-inference.com/likelihood#likelihoodprinciple

[11] Rossi P, Allenby G, McCulloch R. *Bayesian Statistics and Marketing (pp. 4)*. John Wiley & Sons, 2005.

[12] Efron, Bradley. *Controversies in the Foundations of Statistics*. The American Mathematical Monthly, Vol. 85, No. 4 (Apr., 1978), pp. 231-246.

[13] Efron, Bradley. *A 250-year Argument: Belief, Behavior, and the Bootstrap*. Bull. Amer. Math. Soc. 50 (2013), 129-146.

[14] http://www.quora.com/Statistics-academic-discipline/What-is-a-confidence-interval-in-laymans-terms

[15] http://www.quora.com/What-is-the-difference-between-Bayesian-and-frequentist-statisticians

[16] http://en.wikipedia.org/wiki/Confidence_interval#Meaning_and_interpretation

[17] Thomas P. Minka. *Old and New Matrix Algebra Useful for Statistics*. December 28, 2000.

[18] http://en.wikipedia.org/wiki/Matrix_calculus. Accessed on February 7, 2019

[19] S. R. Searle and H. V. Henderson. *A Primer on Differential Calculus for Vectors and Matrices*. BU-1047-MB, 1993.

[20] Steven W. Nydick. *A Different(ial) Way Matrix Derivatives Again*. May 17, 2012.

[21] Steven W. Nydick. *With(out) A Trace Matrix Derivatives the Easy Way*. May 16, 2012.

[22] Sam Roweis. *Matrix Identities*. June 1999.

[23] Terry Tao. *Matrix identities as derivatives of determinant identities*. January 13, 2013

[24] P. G. Harrison. *Sloppy Derivations of Itô's Formula and the Fokker-Planck Equations*. April 2005.

[25] Fabrice Douglas Rouah. *Heuristic Derivation of the Fokker-Planck Equation*. Retrieved 2/2/2019.

[26] James Stewart *Calsulus - Early Transcendentals*. Cengage Learning, 2012

[27] Walter Rudin *Principles of Mathematical Analysis*. McGraw-Hill Companies, Inc., 1976.

[28] H. L. Royden *Real Analysis*. Pearson Eduction, Inc., 1988.

[29] Erwin Kreyszig *Introductory Functional Analysis with Applications*. Wiley, 1989.

[30] Gerald B. Folland *Real Analysis: Modern Techniques and Their Applications*. Wiley, 1999.

[31] Alberto Torchinsky *Real Variables*. Westview Press, 1995.

[32] Joseph A. Gallian *Contemporary Abstract Algebra (7th Edition)*. Cengage Learning, 2010.

[33] James Stewart *Calsulus - Early Transcendentals*. Cengage Learning, 2012

[34] Walter Rudin *Principles of Mathematical Analysis*. McGraw-Hill Companies, Inc., 1976.

[35] H. L. Royden *Real Analysis*. Pearson Eduction, Inc., 1988.

[36] Erwin Kreyszig *Introductory Functional Analysis with Applications*. Wiley, 1989.

[37] Gerald B. Folland *Real Analysis: Modern Techniques and Their Applications*. Wiley, 1999.

[38] Alberto Torchinsky *Real Variables*. Westview Press, 1995.

[39] http://normaldeviate.wordpress.com/2012/11/17/what-is-bayesianfrequentist-inference/

[40] http://www.quora.com/What-is-the-difference-between-Bayesian-and-frequentist-statisticians

[41] http://www.bayesian-inference.com/advantagesbayesian

[42] http://www.bayesian-inference.com/likelihood#likelihoodprinciple

[43] Rossi P, Allenby G, McCulloch R. *Bayesian Statistics and Marketing (pp. 4)*. John Wiley & Sons, 2005.

[44] Efron, Bradley. *Controversies in the Foundations of Statistics*. The American Mathematical Monthly, Vol. 85, No. 4 (Apr., 1978), pp. 231-246.

[45] Efron, Bradley. *A 250-year Argument: Belief, Behavior, and the Bootstrap*. Bull. Amer. Math. Soc. 50 (2013), 129-146.

[46] http://www.quora.com/Statistics-academic-discipline/What-is-a-confidence-interval-in-laymans-terms

[47] http://www.quora.com/What-is-the-difference-between-Bayesian-and-frequentist-statisticians

[48] http://en.wikipedia.org/wiki/Confidence_interval#Meaning_and_interpretation

[49] Thomas P. Minka. *Old and New Matrix Algebra Useful for Statistics*. December 28, 2000.

[50] http://en.wikipedia.org/wiki/Matrix_calculus. Accessed on February 7, 2019

[51] S. R. Searle and H. V. Henderson. *A Primer on Differential Calculus for Vectors and Matrices*. BU-1047-MB, 1993.

[52] Steven W. Nydick. *A Different(ial) Way Matrix Derivatives Again*. May 17, 2012.

[53] Steven W. Nydick. *With(out) A Trace Matrix Derivatives the Easy Way*. May 16, 2012.

[54] Sam Roweis. *Matrix Identities*. June 1999.

[55] Terry Tao. *Matrix identities as derivatives of determinant identities*. January 13, 2013

[56] James Stewart *Calsulus - Early Transcendentals*. Cengage Learning, 2012

[57] Walter Rudin *Principles of Mathematical Analysis*. McGraw-Hill Companies, Inc., 1976.

[58] H. L. Royden *Real Analysis*. Pearson Eduction, Inc., 1988.

[59] Erwin Kreyszig *Introductory Functional Analysis with Applications*. Wiley, 1989.

[60] Gerald B. Folland *Real Analysis: Modern Techniques and Their Applications*. Wiley, 1999.

[61] Alberto Torchinsky *Real Variables*. Westview Press, 1995.

[62] http://normaldeviate.wordpress.com/2012/11/17/what-is-bayesianfrequentist-inference/

[63] http://www.quora.com/What-is-the-difference-between-Bayesian-and-frequentist-stat

[64] http://www.bayesian-inference.com/advantagesbayesian

[65] http://www.bayesian-inference.com/likelihood#likelihoodprinciple

[66] Rossi P, Allenby G, McCulloch R. *Bayesian Statistics and Marketing (pp. 4)*. John Wiley & Sons, 2005.

[67] Efron, Bradley. *Controversies in the Foundations of Statistics*. The American Mathematical Monthly, Vol. 85, No. 4 (Apr., 1978), pp. 231-246.

[68] Efron, Bradley. *A 250-year Argument: Belief, Behavior, and the Bootstrap*. Bull. Amer. Math. Soc. 50 (2013), 129-146.

[69] http://www.quora.com/Statistics-academic-discipline/What-is-a-confidence-interval-in-laymans-terms

[70] http://www.quora.com/What-is-the-difference-between-Bayesian-and-frequentist-stat

[71] http://en.wikipedia.org/wiki/Confidence_interval#Meaning_and_interpretation

[72] Thomas P. Minka. *Old and New Matrix Algebra Useful for Statistics*. December 28, 2000.

[73] http://en.wikipedia.org/wiki/Matrix_calculus. Accessed on February 7, 2019

[74] S. R. Searle and H. V. Henderson. *A Primer on Differential Calculus for Vectors and Matrices*. BU-1047-MB, 1993.

[75] Steven W. Nydick. *A Different(ial) Way Matrix Derivatives Again*. May 17, 2012.

[76] Steven W. Nydick. *With(out) A Trace Matrix Derivatives the Easy Way*. May 16, 2012.

[77] Sam Roweis. *Matrix Identities*. June 1999.

[78] Terry Tao. *Matrix identities as derivatives of determinant identities*. January 13, 2013

[79] P. G. Harrison. *Sloppy Derivations of Itô's Formula and the Fokker-Planck Equations*. April 2005.

[80] Fabrice Douglas Rouah. *Heuristic Derivation of the Fokker-Planck Equation*. Retrieved 2/2/2019.

Index