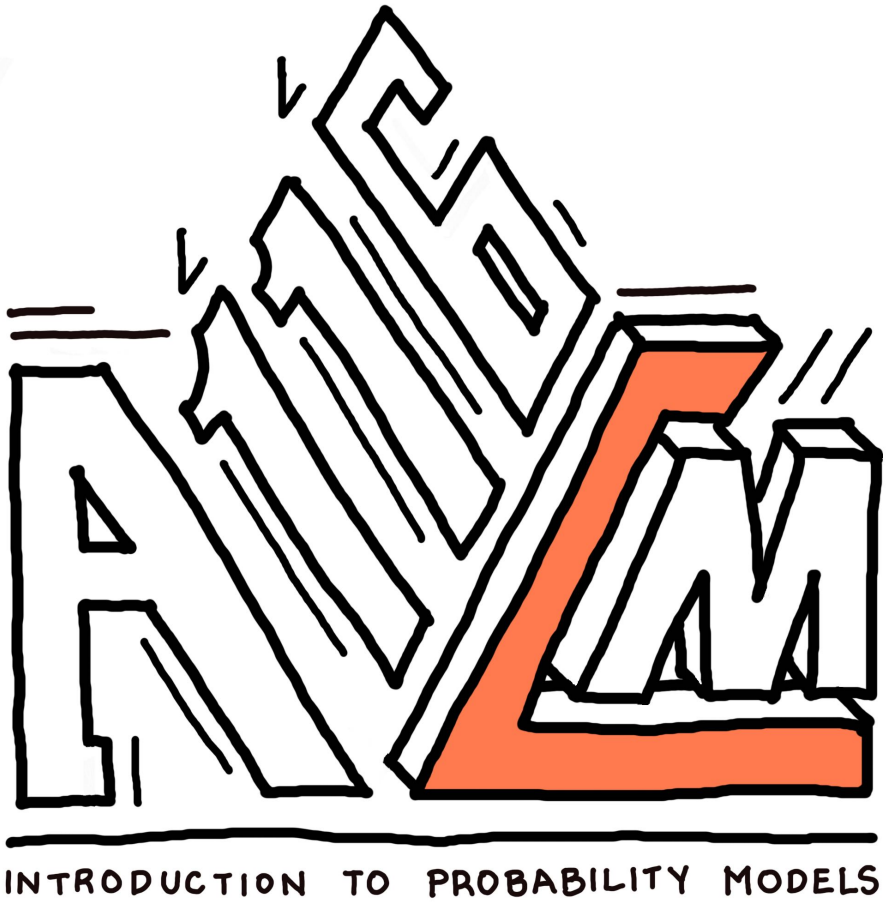
Introduction to Probability Models

ACM 116 / Caltech / Fall 2026

Prof. Kostia Zuev

Notes by Ritvik Teegavarapu



Typeset on November 2, 2026

Copyright ©2023. All rights reserved.

These lecture notes are composed using a template designed by Mathias Legrand with stylistic changes by Joel A. Tropp. This template is licensed under CC BY-NC-SA 3.0.

Contents

Notation and Definitions vi

0 Big Picture 1

1 Probability Models 2

1.1 Probabilistic Models 2

1.2 Basics of Probability 4

2 Random Variables 7

2.1 Cumulative Distribution Functions 8

2.2 Discrete Random Variables 8

2.3 Continuous Random Variables 10

2.4 Expectation 12

2.5 Joint Distribution 13

3 Moment Generating Functions 15

3.1 Independence 15

3.2 Covariance 16

3.3 Moment Generating Functions 18

4 Concentration and Tails 22

4.1 Markov and Chebyshev's Inequality 22

4.2 Law of Large Numbers 25

4.3 Monte Carlo Methods 26

4.4 Central Limit Theorem 27

5 Conditional Probabilities 29

5.1 Motivation 29

5.2 Discrete Setting 30

5.3 Continuous Setting 31

6 Conditional Expectations 33

6.1 Conditional Expectation 33

6.2 Analysis of Quicksort 35

6.3 Compound Random Variables 36

7 More Conditional Probabilities 39
7.1 Indicator Random Variables 39
7.2 Best Prize Problem 41
7.3 Ballot Problem 42
7.4 Double Conditioning 43

8 Random Vectors 45
8.1 Review of Matrix Algebra 45
8.2 Random Vectors and Matrices 47
8.3 Covariance Matrix 48
8.4 Karhunen-Loève Expansion 50
8.5 Transformations of Random Vectors 51

9 Applications of Random Vectors 53
9.1 Wiener Filters 53
9.2 Gaussian Random Vectors 57

10 Introduction to Stochastic Processes 60
10.1 Stochastic Processes 60
10.2 Markov Chains 61
10.3 Counting Processes 62
10.4 Poisson Processes 63

11 Poisson Processes 65
11.1 Limit of Bernoulli Process 65
11.2 Interarrival Times 66
11.3 Waiting Times 68

12 More Poisson Processes 70
12.1 Merging Processes 70
12.2 Splitting Processes 71
12.3 E-Commerce Application 72
12.4 Conditional Arrival Times 73
12.5 Order Statistics 74

13 Applications of Poisson Processes 76
13.1 Insurance Claims 76
13.2 Multi-Type Poisson Process 77
13.3 Estimating HIV Infections 77
13.4 Minimizing Car Encounters 78

14 Brownian Motion 81
14.1 Continuum Limit of Random Walk 81
14.2 Properties of Brownian Motion 84
14.3 Hitting Times and Maximum Values 87
14.4 Variations of Brownian Motion 89

15 General Stochastic Processes 90
15.1 Mean Functions 90
15.2 Covariance and Correlation Functions 91
15.3 Stationarity 93

16 Applications of General Stochastic Processes 96
16.1 Gaussian Processes 96
16.2 Interpreting and Estimating $R_X(\tau)$ 98
16.3 Power Spectral Density 100

Bibliography 102

Notation and Definitions

Set Theory

A set is denoted by a capital letter $\mathsf{A}, \mathsf{B}, \mathsf{C}$. The collection of all subsets of a set S is denoted by $\mathscr{P}(\mathsf{S})$, and the empty set is denoted by $\emptyset$. To build a set, we will use the notation below where $P(x)$ is some relation to be satisfied.

$$\mathsf{B} = \{x \in \mathsf{A} \mid P(x)\}$$

The usual operations of intersection ($\cap$), union ($\cup$), complement (^{c}), and set containment ($\subseteq, \supseteq$) hold as usual. For sequences of numbers, we can index based on an arbitrary set $\mathsf{S} \subseteq \mathbb{N}$ as $(a_i)_{i \in \mathsf{S}}$. For this course, we define $\mathbb{N} = \{1, 2, \dots\}$.

Linear Algebra

The set $\mathbb{M}_{m \times n}$ is used to denote the set of all matrices of that have m rows and n columns. To refer to a column vector (matrix with 1 column), the space $\mathbb{R}^{n \times 1} = \mathbb{R}^n$ is utilized to indicate a n-dimensional vector. For compactness, a column vector will often be written as follows.

$$\mathbf{x} = (x_1, x_2, \cdots, x_n)^\top \in \mathbb{R}^n$$

The determinant of a matrix $A \in \mathbb{M}_{n \times n}$ is denoted by $\det A$, and the transpose by $A^\top$. We will denote $\|\mathbf{x}\|$ is the norm of vector, and the standard Euclidean norm $\|\mathbf{x}\|_2$. The inner product is denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$, which is related to the norm as follows.

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$$

Probability

The function $\mathbb{P}(\cdot) : \mathscr{P}(\mathsf{S}) \to [0, 1]$ will assign every possible subset of the sample space (an event) a probability $p \in [0, 1]$. For shorthand, we will say an event occurs *with probability* (w.p) p.

We will use uppercase italic letters (i.e. X, Y, Z) to denote random variables, and uppercase bold letters (i.e. $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$) to denote random vectors and matrices.

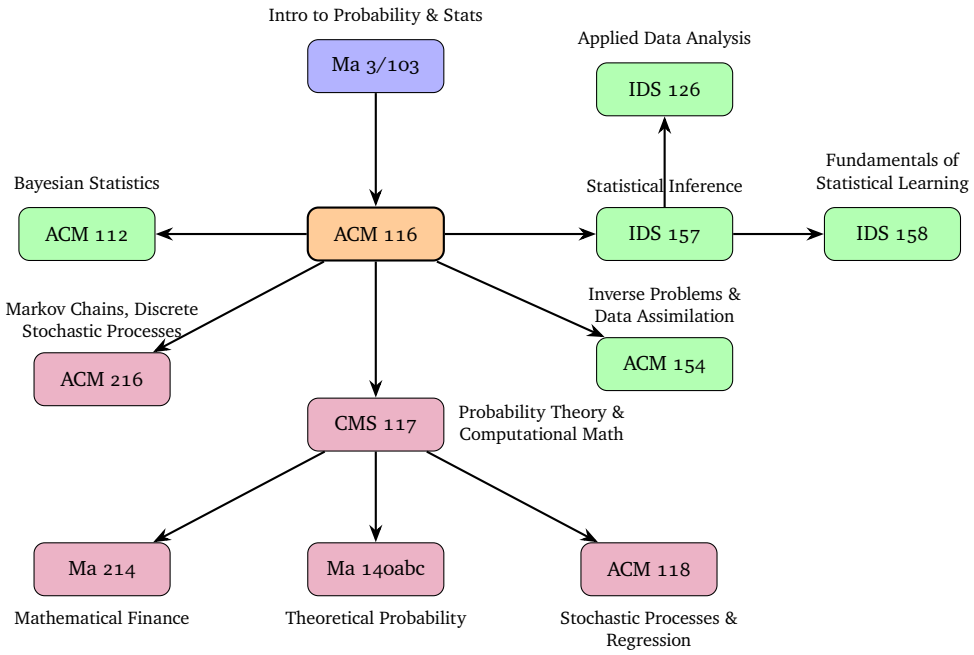
We will use small capital words to indicate named distributions, such as UNIFORM and NORMAL. The symbol $X \sim \cdots$ indicates that X is *distributed* according to $\cdots$. The function F_X will be used to denote the cumulative distribution function (CDF) of X, while f_X will denote the probability density function (PDF) of X.

The operators $\mathbb{E}[X]$ and $\mathbb{V}[X]$ will denote the expectation and variance of a random variable X, respectively. Furthermore, $\mathrm{Cov}(X, Y)$ denotes the covariance between two random variables X and Y.

We will say $X \perp Y$, or that X and Y are *uncorrelated*, if $\mathrm{Cov}(X, Y) = 0$. We will say that $X \perp\!\!\!\perp Y$ if X and Y are *independent*. We say a collection of random variables $\{X_i\}$ is i.i.d if they are independent and identically distributed.

0. Big Picture

Welcome to ACM 116! Probability theory is a fundamental branch of applied mathematics, and is used in all areas of science and engineering. Many courses at Caltech cover elements of probability theory.



The main goals of this course are as follows:

1. Equip science and engineering students with the most important and fundamental probabilistic tools they can use in future studies and research
2. Prepare mathematically oriented students for more advanced probability courses (CMS 117, ACM 118, ACM 216, Ma 144, Ma 214 etc.,)
3. Provide necessary probabilistic background for those who plan to take statistical courses (ACM 112, IDS 126, IDS 157, IDS 158, ACM 154, etc.)

1. Probability Models

We will begin this class by motivating the need to develop probabilistic models for phenomena as opposed to their deterministic counterparts. We will then introduce the fundamental building blocks of probability through the objects of samples spaces and events, and finally discuss some quintessential rules of probability.

Agenda:
1. Probabilistic Models
2. Basics of Probability

1.1 Probabilistic Models

Generally speaking, the course is about *probability models* of natural phenomena. Suppose we are interested in a certain phenomena. If we want to study this phenomena mathematically, we must begin by building a mathematical model which tries to capture the essence.

It is very important to distinguish between the phenomenon (we do not control) and the model (we do control). The model should be ...

1. simple enough to be amenable to analysis (analytically or numerically)
2. complicated enough to be capable of telling us something useful about the phenomenon

There are two general types of models, which are deterministic and probabilistic (stochastic). In **deterministic** models, the output of the model is fully determined by the values of the input parameters. Let's consider a few examples, others being Keplers' and gravitation laws.

Aside: There are other classifications of models (linear vs. nonlinear, or discrete vs. continuous), but these are not of interest to us.

Example 1.1 Suppose we want to measure the area of a lake that for all practical purposes, appears to have a circular shoreline. Then $A = \pi r^2$, where r is the radius, is a deterministic model. It is a model because the shoreline will have some irregularities and not form a perfect circle. It is deterministic because, once the radius is known, the area is known and determined uniquely. ▪

Example 1.2 If we have a circuit with voltage V and resistance R are given for a circuit, then Ohm's law $I = V/R$ predicts the value of the electric current. If we take many electrical circuits with identical voltages and resistances (or even the same circuit), and measure the current, we will get different values for I, but all of them will be very close to V/R with negligible discrepancies. So for all practical purposes, Ohm's law provides a useful deterministic model of reality. ▪

For many phenomena, the deterministic models work well. However, there are also many phenomena which require different models for their investigation.

In **probabilistic** models, fixed values of input parameters lead to an ensemble of different outputs. Input parameters provide some information about the output, but don't determine it exactly as the output remains uncertain. Let's start with a classical example.

Example 1.3 The outcome of a coin toss is clearly not random since given the exact initial conditions, you can use Newtonian mechanics to predict the exact outcome [Kel86]. Coin tossing is a deterministic process, so what is random about it?

If we start from height $h = 0$ and heads up, then the initial vertical speed v and the number of revolutions per second ω fully determine the outcome. So, in principle, we can use a deterministic model developed using Newtonian mechanics $x = f(v, \omega)$. ▪

This model has two drawbacks, the latter of which makes the coin toss look random.

1. It is complicated for such a simple phenomenon.
2. Small changes to the initial conditions lead to different outcomes.

For this model to be useful, we must measure the initial conditions very precisely, which is a non-trivial task. For all practical purposes, a simple probabilistic model is more useful.

$$X = \begin{cases} H & \text{w.p } 1/2 \\ T & \text{w.p } 1/2 \end{cases} \tag{1.1}$$

Here, the input parameter is $p = 1/2$ and the ensemble of outputs (coined the sample space) is $\Omega = \{H, T\}$. It was shown in [Diao7] that real people toss a coin in such a way that it comes up the same way as it started with probability 0.51. In the previous example, we could use a deterministic model at least in principle, but there are many phenomena where this is simply impossible.

Example 1.4 Suppose we want to study arrival times and lengths of service at a medical clinic. This is an important practical problem: its solution helps to figure out how many physicians are needed to serve the facility. There is no hope to build a reasonable deterministic model. We will never be able to predict the exact times. But the good news is that we don't really need the exact times and good probabilistic models exist for this phenomenon. ▪

Other examples include . . .

- predicting the height of a person from the height of their parents
- number of α-particles emitted during a specified period of time by a piece of radioactive material from its shape, mass, and chemical composition
- predicting a future stock price from the previous history

As such, probability models will be the main focus of the course. But before we discuss how to build models and analyze them, we need a certain level of understanding of the most fundamental notions of probability. This will be our goal for the next few lectures. There are two approaches to the study of probability, the former of which will be our principle.

Aside: We assume most of the content was covered in Ma 3, so we will cover it fast.

1. heuristic: attempts to develop an intuitive feel for probability
2. formal: focuses on rigor through the lens of measure theory

Warning 1.5 In particular, there are measure-theoretic subtleties that we will not consider that render some of the theorems we state not *entirely* correct. For example, when saying a statement holds for **all** functions, we must restrict our attention to measurable functions. However, all of the functions in this course are "nice enough" that this is automatically true. ▪

1.2 Basics of Probability

Let us start with a quick review of the basics of probability theory. Suppose we study a phenomenon or experiment whose outcome is uncertain.

Definition 1.6 (Probability Space). The **probability space** $(\Omega, \mathscr{E}, \mathbb{P})$ is defined by the following three objects:

1. **Sample Space**: The set of all possible outcomes of an experiment is called the sample space (Ω).
2. **Event Space**: The set of all possible subsets of Ω is called the event space ($\mathscr{E}$).
3. **Probability Function**: A function $\mathbb{P} : \mathscr{E} \to [0, 1]$ is called a probability distribution

Aside: It is not always possible to assign probability if the sample space is large ($S = \mathbb{R}$). Formally, we define a σ-algebra to limit the subsets, but this is outside the scope of the class.

We say an event $E \in \mathscr{E}$ occurs when the outcome of the experiment $\omega \in E$. To every event $E \in \mathscr{E}$, we want to assign $\mathbb{P}(E)$, which is how likely it is for E to occur. We do not want $\mathbb{P}$ to be arbitrary, so it should satisfy natural properties.

Definition 1.7 (Axioms of Probability). The probability function must satisfy the following properties:

1. $0 \leq \mathbb{P}(A) \leq 1$
2. $\mathbb{P}(S) = 1$
3. $\mathbb{P}(\emptyset) = 0$
4. $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$
5. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
6. $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i)$ for disjoint $\{A_i\}_{i=1}^{\infty}$

Aside: This is not the minimal description of the axioms; in fact, everything can be inferred from {1, 2, 6}.

With this framework, one can now define when one event does not influence how likely another event is to occur.

Definition 1.8 (Independence). Two events A and B are **independent** ($A \perp\!\!\!\perp B$) if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

Independence can arise in two distinct ways, in that we either **assume** that two events are independent (two coin tosses), or **derive** independence of A and B by verifying Definition 1.10.

Example 1.9 Toss a fair dice ($S = \{1, \ldots, 6\}$), and let $A = \{2, 4, 6\}, B = \{1, 2, 3, 4\}$.

$$\mathbb{P}(A \cap B) = \frac{1}{3} = \frac{1}{2} \cdot \frac{2}{3} = \mathbb{P}(A) \cdot \mathbb{P}(B) \tag{1.2}$$

We can verify that A and B are independent. ∎

Definition 1.10 (Joint/Pairwise Independence). A set of events $\{A_1, \ldots, A_n\}$ are **jointly independent** if for every subset $\{A_{i_1}, \ldots, A_{i_m}\}$ where $m \leq n$

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_m}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_m}).$$

A set of events $\{A_1, \ldots, A_n\}$ are **pairwise independent** if for all $i, j \in \{1, \ldots, n\}$,

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(A_j)$$

You can see that pairwise independent does not imply joint independence in the case of three events A, B, C.

Suppose we know that $B \subset S$ has occured. This information changes the probability of occurence of other events. For example, if $B \cap C = \emptyset$, the new probability of C is zero. If $A \subseteq B$, the new probability of A is larger than the old one. This "new" probability is coined the **conditional** probability.

Definition 1.11 (Conditional Probability). If $\mathbb{P}(B) > 0$, then the conditional probability of A given B is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Intuitively, $\mathbb{P}(A \mid B)$ is the fraction of times the event A occurs among those in which B occurs.

Theorem 1.12 (Conditional Probability Properties). The conditional probability obeys the following properties.

- For any fixed B, $\mathbb{P}(\cdot \mid B)$ is a probability distribution on the sample space Ω that satisfies all of the axioms.
- In general, $\mathbb{P}(A \mid B) \neq \mathbb{P}(B \mid A)$.
- If $A \perp\!\!\!\perp B$, then $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.

Aside: The last property tells us that knowledge of B does not inform you of A.

Example 1.13 If $A = \{\text{measles}\}, B = \{\text{spots}\}$, then $\mathbb{P}(B \mid A) = 1$ while $\mathbb{P}(A \mid B) < 1$. ▪

A useful tool for computing probabilities is the **law of total probability**.

Theorem 1.14 (Law of Total Probability). Let $\{B_i\}_{i=1}^n$ be a partition of Ω, meaning that the following properties are satisfied.

1. $\bigcup_{i=1}^n B_i = S$ (jointly exhaustive)
2. $B_i \cap B_j = \emptyset$ (mutually exclusive)
3. $\mathbb{P}(B_i) > 0$

Then, the following relation holds true.

$$\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A \mid B_i) \cdot \mathbb{P}(B_i)$$

If we use Equation 1.11 and flip the events, we can recover Bayes' Formula.

Theorem 1.15 (Bayes' Formula). For events A, B such that $\mathbb{P}(B) > 0$,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}$$

We finish the lecture with a classical example of conditional probability.

Example 1.16 In the Monty Hall game show, there are three doors numbered $\{1, 2, 3\}$. A uniformly random door has a car behind it, while the other two doors each have a goat behind them. The contestants usually prefer cars to goats.

1. The contestant first picks a door.
2. The host, who knows where the car is, opens a uniformly random door that was not picked by the contestant and has a goat behind it.
3. The contestant then chooses to stay with the door she already picked, or to switch to the other remaining door.

What should the contestant do in the last step? ▪

Without loss of generality, we can assume that the contestant initially picked door number 1 and the host opened door number 3. Let C_1, C_2, C_3 be the events that the car is behind door 1, 2 and 3 respectively. We would like to calculate $\mathbb{P}(C_2 \mid H_3)$ and $\mathbb{P}(C_1 \mid H_3)$. We can use Bayes' rule to write:

$$\mathbb{P}(C_2 \mid H_3) = \frac{\mathbb{P}(H_3 \mid C_2) \cdot \mathbb{P}(C_2)}{\mathbb{P}(H_3)}. \tag{1.3}$$

It is clear that $\mathbb{P}(H_3 \mid C_2) = 1$ (as the host cannot reveal what the contestant chose or the car), and $\mathbb{P}(C_2) = 1/3$. To calculate $\mathbb{P}(H_3)$, note that by symmetry we must have $\mathbb{P}(H_3) = \mathbb{P}(H_2)$ and we have $\mathbb{P}(H_3) + \mathbb{P}(H_2) = 1$, so $\mathbb{P}(H_3) = 1/2$. This gives

$$\mathbb{P}(C_2 \mid H_3) = \frac{1/3}{1/2} = \frac{2}{3}. \tag{1.4}$$

Since given H_3, the car is behind either door 1 or door 2, we must have the following.

$$\mathbb{P}(C_1 \mid H_3) + \mathbb{P}(C_2 \mid H_3) = 1 \tag{1.5}$$

Thus, $\mathbb{P}(C_1 \mid H_3) = 1/3$. This proves that the contestant should always choose to switch to the remaining door if she wishes to maximize her probability of winning, even though this is not very intuitive!

References

[Diao7] P. Diaconis. "Dynamics bias in the coin toss". In: (2007).

[Kel86] J. Keller. "The Probability of Heads". In: (1986).

2. Random Variables

In many experiments, we are mostly interested in certain numerical characteristics of outcomes rather than in outcomes. From the previous coin tossing example, we are not interested in the initial velocity and revolution speed of the coin ($\omega = (v, w)$), but rather in whether the coin is heads or tails ($X(\omega) \in \{H, T\}$). These numerical characteristics are called **random variables**.

Agenda:
1. Cumulative Distribution Functions
2. Discrete Random Variables
3. Continuous Random Variables
4. Expectation
5. Joint Distribution

Definition 2.1 (Random Variable). A **random variable** is a function $X : \Omega \to \mathbb{R}$ that assigns to every outcome $\omega \in \Omega$ a realization, which is a real number $x = X(\omega)$.

This terminology is unfortunate since the random variable is neither random (as it is a deterministic function of the sample space) nor a variable (as it is a function). However, this terminology is universally accepted, so we will not deviate from it.

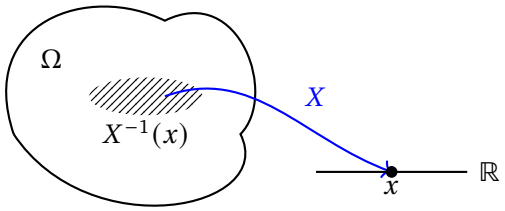
To relate this to the probability distribution over the space $(\Omega, \mathcal{E}, \mathbb{P})$, the random variable induces a *new* probability space over the real numbers $(\mathbb{R}, \mathcal{F}, \mathbb{P}_X)$ where $\mathbb{P}_X$ is defined below.

Definition 2.2 (Distribution of Random Variable). For a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ and random variable X, the **distribution** of X is defined by $\mathbb{P}_X$ as

$$\mathbb{P}_X(X \in A) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\})$$

Aside: We will typically omit the subscript X when the random variable of interest is clear.

This is visualized below where $A = \{x\}$, but can be compactly stated as the probability of the pre-image of the set under the random variable.



While this definition of a random variable and its distribution unify our treatment of the content with the measure-theoretic formulation of probability, the underlying sample space is not particularly important and we will work directly with random variables.

This "covering up" of the probability space is exactly why random variables are called random, since all of the inherent randomness arises from the sample space which we never see. Nonetheless, it is useful to keep in the back of your mind that the sample space is there lurking in the background.

2.1 Cumulative Distribution Functions

Random variables are characterized by their cumulative distribution function.

Definition 2.3 (CDF). The **cumulative distribution function** (CDF) F_X of the random variable X is a function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = \mathbb{P}(X \leq x)$$

A natural question is why this is a useful function to look at. This is because the CDF uniquely determines the random variable. In particular, we can recover $\mathbb{P}(X \in A)$ for any $A \subset \mathbb{R}$ given the CDF.

Theorem 2.4 (Uniqueness of CDF). Let X and Y be random variables with CDFs F_X and F_Y. If $F_Y(x) = F_X(x)$ for all x, then $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for all $A \subset \mathbb{R}$.

Aside: The proof of this result is quite involved, with an argument involving Caratheodory's extension theorem.

The more important insight to extract from this result is that the CDF entirely determines the distribution of X, and all probability questions about X can be answered through the lens of F_X.

We can also discuss relevant properties of the CDF that we expect. The name "Cádlág" is a common name used to denote non-decreasing and right-continuous functions that is used in the theory of stochastic processes.

Theorem 2.5 (Properties of CDF). The CDF $F_X : \mathbb{R} \to [0, 1]$ of a random variable satisfies the following properties.

1. **Non-decreasing**: If $a \leq b$, then $F_X(a) \leq F_X(b)$.
2. **Asymptotics**: $\lim_{x \to \infty} F_X(x) = 1$ and $\lim_{x \to -\infty} F_X(x) = 0$
3. **Right Continuity**: For each $x \in \mathbb{R}$, we have $\lim_{y \to x^+} F_X(y) = F(x)$

Remarkably so, the *converse* is also true in that any function that obeys the above properties is the CDF of a unique random variable. The converse is a technical result that requires measure theory to establish.

Aside: There are random variables that are neither, but both are outside the scope of the class.

2.2 Discrete Random Variables

Random variables are typically classified as either discrete or continuous, which are based on the set of values they take on.

Definition 2.6 (Discrete Random Variable). A random variable is called **discrete** if it takes a finite or a countable number of possible values $\{x_1, x_2, \dots\}$.

It is convenient to describe a discrete random variable in terms of its probability mass function (PMF).

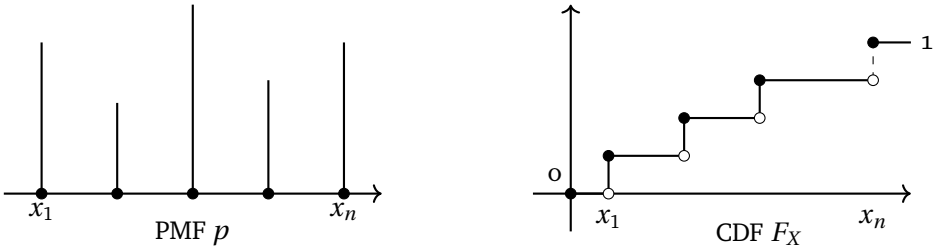
Definition 2.7 (PMF). The **PMF** of a discrete random variable X with values in $\{x_1, x_2, \dots\}$ is defined by a function p such that

$$p(x_i) = \mathbb{P}(X = x_i)$$

It is clear that $\sum_i p(x_i) = 1$ since X must take one of those values. In particular, we can define the CDF as follows.

$$F(x) = \mathbb{P}(X \leq x) = \sum_{i:x_i \leq x} p(x_i) \tag{2.1}$$

We can see this visually that the PMF has point masses at the points $\{x_1, \ldots, x_n\}$, and the CDF has all of the properties we described above.



Remark 2.8 There is a well-known, old dichotomy in physics: is our world discrete or continuous? In general relativity, the space-time is continuous. But GR does not take quantum effects into account. A theory (that does not exist yet) that tries to bridge GR with quantum mechanics is called quantum gravity. There are several approaches to Q.G. One is string theory, where the space-time is still continuous. But there are other candidates: loop Q.G., causal sets, etc., that assume that space-time is discrete. Probability is one of the most "physical" branches of math. It also has this discrete vs continuous dichotomy.

Bernoulli Random Variable

Suppose we are interested in an experiment, whose outcome can be classified as either a "success" or as a "failure." Let $X = 1$ in the case of success and $X = 0$ in the case of failure. Then the PMF for $X \sim \text{Bernoulli}(p)$ is defined as follows for $k \in \{0, 1\}$.

$$\begin{cases} p(0) = \mathbb{P}(X = 0) = 1 - p \\ p(1) = \mathbb{P}(X = 1) = p \end{cases} \tag{2.2}$$

Binomial Random Variable

Suppose we perform n independent experiments, each of which results in a success with probability p and in a failure with probability $(1 - p)$. Let X be the number of successes in n independent experiments. Then the PMF for $X \sim \text{Binomial}(n, p)$ is defined as follows for $k \in \{0, 1, \ldots, n\}$.

$$p(k) = \binom{n}{k} p^k (1 - p)^{n-k} \tag{2.3}$$

Geometric Random Variable

As before, we perform independent experiments, each of which results in a success with probability p, but continue going *until* a success occurs. Then the PMF for $X \sim \text{Geometric}(p)$ is defined as follows for $k \in \mathbb{N}$.

$$p(k) = p(1 - p)^{k-1} \tag{2.4}$$

Poisson Random Variable

Let us consider the setup of the binomial random variable in that we perform n independent experiments with success probability p. A natural question is what happens in the limiting regime of $n \to \infty$?

For a fixed p, this question is not interesting since $X \to \infty$ in this limit. However, if we take $p \to 0$ simultaneously in a consistent manner (i.e. $np = \lambda$), we can deduce a limiting distribution.

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \tag{2.5}$$

$$= \frac{n(n-1)\dots(n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^k} \tag{2.6}$$

$$\lim_{n\to\infty, p\to 0} \binom{n}{k} p^k (1-p)^{n-k} = \frac{e^{-\lambda} \cdot \lambda^k}{k!} \tag{2.7}$$

Then the PMF for $X \sim \text{Poisson}(\lambda)$ is defined as follows for $k \in \mathbb{N} \cup \{0\}$.

$$p(k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} \tag{2.8}$$

An example will clarify how the Poisson distribution inherits the intuition from the binomial distribution, which is called the *law of rare events*.

Example 2.9 We view cars on the highway for some time T. The expected rate at which you see a car is every λ minutes. If you model whether a car is seen in an infinitesimal time interval of length Δt as $X \sim \text{Bernoulli}(\lambda \Delta t)$, the total number of occurences will be modeled by $S \sim \text{Binomial}(T/\Delta t, \lambda \Delta t)$. This will recover the Poisson distribution as we take the limit of $\Delta t \to 0$. ∎

2.3 Continuous Random Variables

To define a random variable, we need to specify how to compute the probabilities $\mathbb{P}(X \in A)$ for all $A \subset \mathbb{R}$. In the discrete setting, we can do this through the PMF.

$$\mathbb{P}(X \in A) = \sum_{x_i \in A} p(x_i) \tag{2.9}$$

As often in mathematics, to define a continuous analog of a discrete concept, we must replace $\sum$ with $\int$ with appropriate modifications.

Aside: There is a distinction between continuous random variables and *absolutely* continuous random variables (which we discuss here). However, it is not of importance in this class.

Definition 2.10 (Continuous Random Variable). A random variable X is **continuous** if there exists a non-negative function $f : \mathbb{R} \to \mathbb{R}^{\geq 0}$ called the probability density function (PDF) with the following property for any subset $A \subset \mathbb{R}$

$$\mathbb{P}(X \in A) = \int_A f_X(x)\, \mathrm{d}x$$

The following are properties of the PDF that we can deduce from the previous definitions.

Property 2.11

$$\int_{-\infty}^{\infty} f_X(x)\, \mathrm{d}x = \mathbb{P}(X \in (-\infty, \infty)) = 1$$

Property 2.12

$$\int_a^b f_X(x)\,\mathrm{d}x = \mathbb{P}(X \in [a,b]) \implies \mathbb{P}(X = a) = 0$$

Property 2.13

$$\frac{\mathrm{d}F_X(x)}{\mathrm{d}x} = f_X(x)$$

We can use the PDF to answer all probability questions about X. However, unlike the PMF, it is not true that $f_X(a) = \mathbb{P}(X = a)$. To obtain an intuitive interpretation of the PDF, consider the ϵ-neighborhood of x.

$$\mathbb{P}\left(x - \frac{\epsilon}{2} \le X \le x + \frac{\epsilon}{2}\right) = \int_{x-\epsilon/2}^{x+\epsilon/2} f_X(x)\,\mathrm{d}x \approx \epsilon \cdot f_X(x) \tag{2.10}$$

Thus, $f_X(x)$ is a measure of how likely it is that the random variable X will be near x. When the underlying random variable is clear, we will omit the subscript for brevity.

Uniform Random Variable

Suppose we are interested in generating a random real number on the interval $[\alpha, \beta]$ where any real number is equally likely. Then the PDF of the random variable $X \sim \textsc{Uniform}[\alpha, \beta]$ is defined as follows.

$$f_X(x) = \begin{cases} \frac{1}{\beta - \alpha} & x \in (\alpha, \beta) \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

Exponential Random Variable

Suppose we want to model the lifetime of a light-bulb whose expected lifetime is λ^{-1} through random variable X. Then the PDF of the random variable $X \sim \textsc{Exponential}(\lambda)$ is defined as follows for $x \in \mathbb{R}^{\ge 0}$.

$$f_X(x) = \lambda e^{-\lambda x} \tag{2.12}$$

It is also used to model the times between rare events.

Normal/Gaussian Random Variable

We say that $X \sim \textsc{Normal}(\mu, \sigma^2)$ is a normal random variable with parameters μ and σ^2 with PDF defined as follows.

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \tag{2.13}$$

At first glance, the PDF is quite involved and does not "normal." However, the normal PDF is the single most important distribution in probability and statistics as many phenomena in nature can be accurately modeled by sums of random variables which can be approximated with the normal distribution with the central limit theorem.

We say that $X \sim \textsc{Normal}(0, 1)$ is the standard normal distribution, with PDF $\varphi(x)$ and CDF $\Phi(x)$. The normal distribution is convenient in that we can rescale to get the standard normal through a linear transformation.

$$X \sim \textsc{Normal}(\mu, \sigma^2) \implies Z = \frac{X - \mu}{\sigma} \sim \textsc{Normal}(0, 1) \tag{2.14}$$

Remark 2.14 Vladamir Arnold's principle states that if a notion bears a personal name, then this name is not the name of the discoverer. The Arnold Principle is applicable to itself as well as to the Gaussian distribution: the standard normal curve was discovered around 1720 by Abraham de Moivre.

2.4 Expectation

The expectation of a random variable X is the average value of X.

Definition 2.15 (Expected Value). The **expected value** of a random variable X is

$$\mathbb{E}[X] = \begin{cases} \sum_i x_i \cdot p(x_i) & \text{discrete } X \\ \int_{-\infty}^{\infty} x f(x)\, \mathrm{d}x & \text{continuous } X \end{cases}$$

Aside: The expectation is linear, in that $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$.

One can think of $\mathbb{E}[X]$ as the average value you would obtain if you computed the numerical average of $\frac{1}{n}\sum_i x_i$ where $x_i = X(\omega_i)$ for $i = 1, \ldots, n$ of a common random variable X. The fact that $\mathbb{E}[X] \approx \frac{1}{n}\sum x_i$ is a theorem called the law of large numbers.

Another intuitive interpretation is physical, in that the expected value is the center of mass. Naturally, the spread of mass can be computed as follows.

Definition 2.16 (Variance). The **variance** of X is $\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

By expanding the expectation, we can simplify this to a common formula used for computations of the variance.

$$\mathbb{V}[X] = \mathbb{E}[X^2 - 2\mathbb{E}[X] \cdot X + \mathbb{E}[X]^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{2.15}$$

Expectations of the form $\mathbb{E}[X^n]$ have a special name known as **moments**, which are different one-number summaries of the distribution than the expectation.

Aside: The moment may be infinite for a random variable that does not "go to 0" fast enough.

Definition 2.17 The n-th **moment** of a random variable X is

$$\mathbb{E}[X^n] = \begin{cases} \sum_i x_i^n \cdot p(x_i) & \text{discrete } X \\ \int_{-\infty}^{\infty} x^n f(x)\, \mathrm{d}x & \text{continuous } X \end{cases}$$

The natural extension of this if we can compute $\mathbb{E}[g(X)]$ for an arbitrary function g. If we let $Y = g(X)$, then Y is random due to the randomness of X. However, the PDF f_Y may not be easily written in terms of f_X. Luckily, we can be lazy!

Theorem 2.18 (LOTUS). For any random variable X with CDF F_X, we can write the Riemann-Stieltjes integral as follows.

$$\mathbb{E}[g(X)] = \int g(x)\, \mathrm{d}F_X(x)$$

Aside: LOTUS stands for "law of the unconscious statistician."

The Riemann-Stieltjes integral allows us to unify the formulae for the discrete and continuous random variables as we show below.

$$\mathbb{E}[g(X)] = \int g(x)\, \mathrm{d}F_X(x) = \int g(x) \cdot f_X(x)\, \mathrm{d}x \qquad \text{(continuous } X\text{)} \tag{2.16}$$

$$\mathbb{E}[g(X)] = \int g(x)\, \mathrm{d}\left(\sum_i p(x_i) \cdot H(x - x_i)\right) \qquad \text{(discrete } X\text{)} \tag{2.17}$$

$$= \sum_i p(x_i) \int g(x) \cdot \delta(x - x_i)\, \mathrm{d}x \qquad \text{(interchange sum)} \tag{2.18}$$

$$= \sum_i p(x_i) \cdot g(x_i) \qquad \text{(sifting property)} \tag{2.19}$$

2.5 Joint Distribution

In many applications, one random variable is not enough for modeling the phenomenon of interest, and we need to consider several random variables jointly. These random variables ca be described in terms of their **joint** CDF.

Definition 2.19 (Joint CDF). The **joint CDF** of random variables X and Y is

$$F_{X,Y}(x,y) = \mathbb{P}(X \le x, Y \le y)$$

It is simple to obtain the *marginal* CDFs of X and Y from the joint CDF.

$$F_X(x) = \mathbb{P}(X \le x) = \mathbb{P}(X \le x, Y < \infty) = F_{X,Y}(x,\infty) \tag{2.20}$$
$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X < \infty, Y \le y) = F_{X,Y}(\infty,y) \tag{2.21}$$

In the discrete setting, we can define the joint PMF where $\{x_i\}, \{y_j\}$ are the values X, Y take on, respectively.

Definition 2.20 (Joint PMF). The **joint PMF** of random variables X and Y is

$$p_{X,Y}(x_i,y_j) = \mathbb{P}(X = x_i, Y = y_j)$$

The *marginal* PMFs of X and Y are computed from the joint PMF.

$$p_X(x_i) = \mathbb{P}(X = x_i) = \sum\nolimits_j p(x_i,y_j) \tag{2.22}$$
$$p_Y(y_j) = \mathbb{P}(Y = y_j) = \sum\nolimits_i p(x_i,y_j) \tag{2.23}$$

In the continuous setting, we define joint continuity of two random variables if they admit a density.

Definition 2.21 (Joint Continuity). We say that X, Y are **jointly continuous** if there exists a non-negative function $f_{X,Y}(x,y)$ (the joint PDF of X, Y) such that

$$\mathbb{P}(X \in A, Y \in B) = \int_B \int_A f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

Once again, the marginal PDFs of X and Y can be computed from the joint PDF in a nearly identical computation to the discrete setting.

$$\mathbb{P}(X \in A) = \mathbb{P}(X \in A, Y \in (-\infty,\infty)) \tag{2.24}$$
$$= \int_{-\infty}^{\infty} \int_A f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \tag{2.25}$$
$$= \int_A \left[\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y \right] \mathrm{d}x \tag{2.26}$$

By definition of the PDF, we then have the following definition of the marginals.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y \tag{2.27}$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x \tag{2.28}$$

From the setting of one random variable, we expect to have some differential relation between the CDF and PDF, which is indeed true.

$$F_{X,Y}(x,y) = \mathbb{P}(X \le x, Y \le y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \tag{2.29}$$

$$\frac{\partial F_{X,Y}}{\partial x \partial y} = f_{X,Y}(x,y) \tag{2.30}$$

Linearity of expectation holds for multiple random variables, which we derive in the setting of two random variables. Note that the joint distribution can be defined for n random variables in an analogous way.

$$\mathbb{E}[\alpha X + \beta Y] = \iint (\alpha x + \beta y) f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \tag{2.31}$$

$$= \alpha \iint x \cdot f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y + \beta \iint y \cdot f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \tag{2.32}$$

$$= \alpha \int x \left[\int f_{X,Y}(x,y)\,\mathrm{d}y\right]\mathrm{d}x + \beta \int y \left[\int f_{X,Y}(x,y)\,\mathrm{d}x\right]\mathrm{d}y \tag{2.33}$$

$$= \alpha \int x f_X(x)\,\mathrm{d}x + \beta \int y f_Y(y)\,\mathrm{d}y \tag{2.34}$$

$$= \alpha \cdot \mathbb{E}[X] + \beta \cdot \mathbb{E}[Y] \tag{2.35}$$

More generally, we have the following result.

Theorem 2.22 (Linearity of Expectation). For random variables $\{X_i\}_{i=1}^n$ and scalars $\{\alpha_i\}_{i=1}^n$, the following is true.

$$\mathbb{E}\left[\sum_{i=1}^n \alpha_i X_i\right] = \sum_{i=1}^n \alpha_i \cdot \mathbb{E}[X_i]$$

3. Moment Generating Functions

Just as we defined the concept of independence for events, we can define independence for random variables.

Agenda:
1. Independence
2. Covariance
3. Moment-Generating Functions

3.1 Independence

Intuitively, X and Y are independent if the outcome of X in no way influences the outcome of Y. In other words, any information about X does not tell us anything about Y and vice versa. The formal definition is the following.

Definition 3.1 (Independence). Two random variables X and Y are **independent** ($X \perp\!\!\!\perp Y$) if for all subsets $A, B \subset \mathbb{R}$, the following is true.

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(X \in B)$$

In principle, to check whether X and Y are independent, we would need to check the definition for all subsets A, B. Fortunately, we have the following equivalences.

Theorem 3.2 (Independence Relations). If $X \perp\!\!\!\perp Y$, then the following hold true.

1. $F_{X,Y}(x, y) = F_X(x) \cdot F_y(y)$
2. $p_{X,Y}(x, y) = p_X(X) \cdot p_Y(y)$ (X, Y discrete)
3. $f_{X,Y}(x, y) = f_X(X) \cdot f_Y(y)$ (X, Y continuous)

Independence is a helpful tool in computing the expectations of complicated expressions.

Aside: Formally, these functions must be bounded and measurable.

Theorem 3.3 (Independence Relations). The random variables X, Y are independent ($X \perp\!\!\!\perp Y$) if and only if the following holds true for any function f, g.

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)]$$

Independence is a highly unintuitive concept. To see this at play, consider the following example.

Example 3.4 Let $X = \text{Uniform}\{1, 2, \ldots, 100\}$, defining events $A = \{X \text{ divisible by } 2\}$, and $B = \{X \text{ divisible by } 5\}$. We can see that

$$\mathbb{P}(X \in A \cap B) = \frac{1}{10} = \frac{1}{2} \cdot \frac{1}{5} = \mathbb{P}(X \in A) \cdot \mathbb{P}(X \in B) \tag{3.1}$$

However, if $X = \text{Uniform}\{1, 2, \ldots, 101\}$, we can see that the independence no longer holds.

$$\mathbb{P}(X \in A \cap B) = \frac{10}{101} \neq \frac{1}{2} \cdot \frac{1}{5} = \mathbb{P}(X \in A) \cdot \mathbb{P}(X \in B) \tag{3.2}$$

∎

3.2 Covariance

There is no unique natural measure of degree of dependence between random variables, with different measures understanding different aspects. One of these measures is the **covariance**.

Definition 3.5 (Covariance). The covariance between X and Y is

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])]$$

An alternative formula is often more convenient.

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y] \tag{3.3}$$

If $X = Y$, then $\text{Cov}(X, X) = \mathbb{V}[X]$.

An important result to note is that $X \perp\!\!\!\perp Y$ implies $\text{Cov}(X, Y) = 0$, which we say is uncorrelated ($X \perp Y$). However, the converse is not true.

Example 3.6 If $X \sim \text{Uniform}[-1, 1]$ and $Y = X^2$, it is clear that X and Y are dependent. However, they are uncorrelated.

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[X \cdot X^2] - \mathbb{E}[X] \cdot \mathbb{E}[X^2] && (3.4)\\ &= 0 - 0 \cdot \mathbb{E}[X^2] = 0 && (3.5) \end{aligned}$$

▪

Thus, independence is a stronger property than un-correlation. The intuitive explanation of correlation will make this clearer. Covariance is a measure of how much two random variables change together. It can be shown that if $\text{Cov}(X, Y) > 0$, then Y tends to increase as X does. If $\text{Cov}(X, Y) < 0$, then Y tends to decrease as X increases.

Example 3.7 Define events $X = \mathbb{1}_A, Y = \mathbb{1}_B$. We can compute the correlation between the two random variables as follows.

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y] && (3.6)\\ &= \mathbb{P}(X = 1, Y = 1) - \mathbb{P}(X = 1) \cdot \mathbb{P}(Y = 1) && (3.7) \end{aligned}$$

If $\text{Cov}(X, Y) > 0$, then we can use Baye's rule to simplify.

$$\begin{aligned} \mathbb{P}(X = 1, Y = 1) &> \mathbb{P}(X = 1) \cdot \mathbb{P}(Y = 1) && (3.8)\\ \mathbb{P}(Y = 1 \mid X = 1) &> \mathbb{P}(Y = 1) && (3.9) \end{aligned}$$

Therefore, knowledge of $X = 1$ informs you whether $Y = 1$. ▪

There is a rich connection between probability theory and linear algebra that explains the notation $\perp$. We recall the definition of the inner product.

Definition 3.8 (Inner Product). An inner product on a vector space V is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies the following properties.

1. (Bilinearity) $\langle \alpha v_1 + \beta v_2, u \rangle = \alpha \langle v_1, u \rangle + \beta \langle v_2, u \rangle$
2. (Symmetry) $\langle u, v \rangle = \langle v, u \rangle$
3. (Positive-Definiteness) $\langle v, v \rangle \geq 0$ and $\langle v, v \rangle = 0 \iff v = 0$

Consider now a vector space of all random variables with zero mean (centered) and finite variance. This can then be equipped with an inner product as follows, meaning orthogonality is equivalent to two uncorrelated random variables.

$$\langle X, Y \rangle = \text{Cov}(X, Y) = \mathbb{E}[XY] \tag{3.10}$$

Aside: The random variables must be centered to have the positive-definiteness preserved.

This explains the symmetry property of the covariance, but we get two more properties for free by definition of the inner product. Note that these hold for any random variables.

$$\text{Cov}(\alpha X, Y) = \alpha \cdot \text{Cov}(X, Y) \tag{3.11}$$
$$\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z) \tag{3.12}$$

These properties allow us to generalize this for a sum of many random variables, noting that the covariance is linear in both arguments.

$$\text{Cov}\left(\sum_{i=1}^n \alpha_i X_i, \sum_{j=1}^m \beta_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j \text{Cov}(X_i, Y_j) \tag{3.13}$$

With this, we can now understand the variance of the sum, for which we know $\mathbb{E}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \mathbb{E}[X_i]$.

$$\mathbb{V}\left[\sum_{i=1}^n X_i\right] = \text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j\right) \tag{3.14}$$
$$= \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(X_i, X_j) \qquad \text{(linearity)} \tag{3.15}$$
$$= \sum_{i=1}^n \mathbb{V}[X_i] + \sum_{i \neq j} \text{Cov}(X_i, X_j) \tag{3.16}$$
$$= \sum_{i=1}^n \mathbb{V}[X_i] + 2 \cdot \sum_{i>j} \text{Cov}(X_i, X_j) \qquad \text{(symmetry)} \tag{3.17}$$

Therefore, if the random variables are pairwise independent, we then have that the variance passes through the summation.

$$\mathbb{V}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{V}[X_i] \tag{3.18}$$

The concept of independence can be extended to several random variables.

Definition 3.9 (Multi-Variable Independence). Random variables $X_1, \ldots, X_n$ are said to be jointly independent if this holds for all sets $A_1, \ldots, A_n \subset \mathbb{R}$.

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n)$$

There is an analogous criterion of independence in terms of joint and marginal CDFs.

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n) \tag{3.19}$$

This reduces to the corresponding criterion in Theorem 3.2 in terms of PDFs and PMFs for continuous and discrete random variables.

One can also define conditional independence, which means that given knowledge of an event C, the two events A and B are **conditionally independent** if

$$\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C) \cdot \mathbb{P}(B \mid C). \tag{3.20}$$

Aside: Independence and conditional independence do not imply one another.

3.3 Moment Generating Functions

The moment generating function (MGF) is a powerful tool for studying various aspects of probability distributions. It is especially useful in studying sums of independent random variables. In essence, the MGF is an alternative way to specify a random variable (alternative to the CDF, PDF, PMF).

Definition 3.10 (MGF). The moment generating function (MGF) $M_X(t)$ of a random variable X is defined as the following.

$$M_X(t) = \mathbb{E}[\exp(tX)]$$

It is clear that $M_X(0) = 1$ always, but the MGF may not exist for all $t \in \mathbb{R}$. For technical reasons, we will assume that the random variables X are such that $M_X(t)$ is defined in an open neighborhood of 0.

As the name suggests, $M_X(t)$ allows us to extract the polynomial moments of a random variable. This can be seen through a Taylor series expansion of $\exp(tX)$ and differentiating.

$$M_X(t) = \mathbb{E}\left[\sum_{i=1}^{\infty} \frac{t^i X^i}{i!}\right] \tag{3.21}$$

$$= \sum_{i=1}^{\infty} \frac{t^i \cdot \mathbb{E}[X^i]}{i!} \tag{3.22}$$

$$\left.\frac{d^n M_X(t)}{dt^n}\right|_{t=0} = \mathbb{E}[X^n] + \left.\sum_{i=n+1}^{\infty} \frac{t^{i-n} \cdot \mathbb{E}[X^i]}{(i-n)!}\right|_{t=0} \tag{3.23}$$

$$= \mathbb{E}[X^n] \tag{3.24}$$

Aside: Passage of the expectation through an infinite sum generally takes some care.

Without the MGF, we would need integration to find the moments. However, knowledge of the MGF allows to instead utilize differentiation.

Example 3.11 We compute the MGF of $X \sim \text{Poisson}(\lambda)$.

$$M_X(t) = \mathbb{E}[\exp(tX)] \tag{3.25}$$

$$= \sum_{k=0}^{\infty} \frac{e^{kt} e^{-\lambda} \lambda^k}{k!} \tag{3.26}$$

$$= e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\left(e^t \lambda\right)^k}{k!} \tag{3.27}$$

$$= e^{-\lambda} \cdot e^{\lambda e^t} = e^{\lambda(e^t - 1)} \tag{3.28}$$

$$M_X'(0) = \lambda e^t \cdot e^{\lambda(e^t-1)}\Big|_{t=0} \tag{3.29}$$

$$= \lambda = \mathbb{E}[X] \tag{3.30}$$

∎

There are two fundamental properties of MGFs that are often used in applications.

Property 3.12 If $X \perp\!\!\!\perp Y$, then

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) \tag{3.31}$$

Proof. We use the properties of independence to show this.

$$M_{X+Y}(t) = \mathbb{E}[\exp(t(X+Y))] \tag{3.32}$$
$$= \mathbb{E}[\exp(tX) \cdot \exp(tY)] \tag{3.33}$$
$$= \mathbb{E}[\exp(tX)] \cdot \mathbb{E}[\exp(tY)] \tag{3.34}$$
$$= M_X(t) \cdot M_Y(t) \tag{3.35}$$

■

Property 3.13 If $M_X(t) = M_Y(t)$ for t in some neighborhood of zero, then $F_X(t) = F_Y(t)$.

Proof. We prove this in the setting where X, Y are integer-valued random variables on the set $S = \{0, 1, \ldots, n\}$. The equality of the MGFs would imply the following.

$$\sum_{s\in S} e^{ts} \cdot p_X(s) = \sum_{s\in S} e^{ts} \cdot p_Y(s) \tag{3.36}$$
$$\sum_{s\in S} e^{ts} \cdot (p_X(s) - p_Y(s)) = 0 \tag{3.37}$$

Since $\{e^{ts}\}_{s\in S}$ are linearly independent, the above relation can only hold if $p_X(s) = p_Y(s)$, which in turn implies equality of the CDFs.

This justification in the most general setting is outside the scope of the class, but roughly boils down to proving that one can analytically extend the MGF and appeal to a proof through the characteristic function. ■

Remark 3.14 If $M_X = M_Y$ for t in some neighborhood of zero, then the moments are the same, which follows from Property 3.13. However, the converse is not true, in that the moments being the same does not imply that the MGFs are the same. A typical example is the log-normal distribution, for which a distinct distribution has the same moments.

With the aforementioned properties, one can determine the distributions of sums of independent random variables.

Example 3.15 Let $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ be independent random variables. The distribution of $X + Y$ can be computed without MGFs tediously, but quite easily with their help.

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) \tag{3.38}$$
$$= e^{\lambda_1(e^t-1)} \cdot e^{\lambda_2(e^t-1)} \tag{3.39}$$
$$= e^{(\lambda_1+\lambda_2)(e^t-1)} \tag{3.40}$$
$$= M_Z(t) \qquad (Z \sim \text{Poisson}(\lambda_1 + \lambda_2)) \tag{3.41}$$

Therefore, $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$. ▪

For another example, we look at how the MGFs allow us to weaken previous assumptions on limiting behavior of the binomial.

Example 3.16 We introduced the Poisson distribution as a limiting case of the Binomial distribution in the limit as $n \to \infty$.

$$\text{Binomial}(n, \lambda/n) \to \text{Poisson}(\lambda) \tag{3.42}$$

Therefore, for $n \gg 1$ and $p \ll 1$, we have that $\text{Binomial}(n,p) \approx \text{Poisson}(np)$. This can be formulated in terms of an independent sum relation by letting $X = X_1 + \cdots + X_n$ where $X_i \sim \text{Bernoulli}(p)$ i.i.d. This is equivalent to $X \sim \text{Binomial}(n,p)$.

We would like to see if we can extend this approximation to the case where $X_i \sim \text{Bernoulli}(p_i)$ independent, where the p_i may be distinct. We cannot directly apply the above since X is no longer binomially distributed. We can leverage the MGF to show that the approximation in this setting is still sound.

$$M_X(t) = M_{X_1}(t) \cdots M_{X_n}(t) \tag{3.43}$$
$$M_{X_i}(t) = \mathbb{E}[\exp(tX_i)] \tag{3.44}$$
$$= \exp(t)p_i + (1-p_i) \cdot \exp(0) \tag{3.45}$$
$$= 1 + p_i(e^t - 1) \tag{3.46}$$
$$= \exp(p_i(e^t - 1)) \qquad (1 + x \approx e^x \text{ for } x \ll 1) \tag{3.47}$$
$$M_X(t) = \exp\left((e^t - 1) \cdot \sum_{i=1}^n p_i\right) \tag{3.48}$$

Thus, X is approximately distributed to $\text{Poisson}(\sum_{i=1}^n p_i)$ for the $n \gg 1, p \ll 1$ regime. Moreover, one can weaken the independency requirement of the X_i. ▪

The MGF can also be defined for several random variables.

Definition 3.17 The **joint** MGF of $\mathbf{X} = (X_1, \ldots, X_n)$ is as follows, where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product.

$$M_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}[\exp(\langle \mathbf{t}, \mathbf{X} \rangle)]$$

The same benefits are enjoyed for the joint MGF.

Property 3.18 The random variables $\mathbf{X} = (X_1, \ldots, X_n)$ are independent if and only if

$$M_{\mathbf{X}}(\mathbf{t}) = \prod_{i=1}^n M_{X_i}(t_i)$$

Property 3.19 The MGF $M_{\mathbf{X}}(\mathbf{t})$ uniquely determines the CDF $F(x_1, \ldots, x_n)$.

Example 3.20 The multivariate normal distribution is defined where $Z_1, \ldots, Z_n \sim \text{Normal}(0,1)$ i.i.d, and $X_1, \ldots, X_m$ defined as follows.

$$X_1 = \sum_{i=1}^n a_{1i} Z_i + \mu_1 \tag{3.49}$$
$$\vdots \tag{3.50}$$
$$X_m = \sum_{i=1}^n a_{mi} Z_i + \mu_m \tag{3.51}$$

We say that $(X_1, \ldots, X_m)$ are **jointly normally** distributed, or the random vector $\mathbf{X}$ has a multivariate normal distribution if there exists a matrix $A \in \mathbb{M}_{m \times n}$ and $\mu \in \mathbb{R}^m$ such that the following holds true.

$$\mathbf{X} = A\mathbf{Z} + \mu \tag{3.52}$$

We can show that the sum of normal random variables remain normally distributed, which can be shown through the MGF of $X \sim \text{Normal}(\mu, \sigma^2)$ below.

$$M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right) \tag{3.53}$$

From this, it is clear that each of the X_i are normally distributed through the fact above.

$$\mathbb{E}[X_i] = \mu_i \tag{3.54}$$
$$\mathbb{V}[X_i] = \sum_{j=1}^{n} a_{ij}^2 \tag{3.55}$$

We can now compute the joint MGF using the above ideas.

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) &= \mathbb{E}[\exp(t_1 X_1 + \cdots + t_m X_m)] && && (3.56)\\ &= \mathbb{E}[\exp(Y)] && (Y = t_1 X_1 + \cdots + t_m X_m) && (3.57)\\ &= \exp\left(\mu_Y + \frac{\sigma_Y^2}{2}\right) && (Y \text{ normal}) && (3.58) \end{aligned}$$

We need to compute the mean and variance of Y as follows.

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}\left[t_1 X_1 + \cdots + t_m X_m\right] && (3.59)\\ &= \sum_{i=1}^{m} t_i \mu_i && (3.60)\\ \mathbb{V}[Y] &= \mathbb{V}\left[t_1 X_1 + \cdots + t_m X_m\right] && (3.61)\\ &= \sum_{i=1}^{m} t_i^2 \cdot \mathbb{V}[X_i] + 2 \sum_{i>j} t_i t_j \cdot \mathrm{Cov}(X_i, X_j) && (3.62) \end{aligned}$$

Thus, the joint MGF of $\mathbf{X} = (X_1, \ldots, X_n)$ is defined as follows.

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\sum_{i=1}^{m} t_i \mu_i + \sum_{i=1}^{m} \frac{t_i^2}{2} \cdot \mathbb{V}[X_i] + \sum_{i>j} t_i t_j \cdot \mathrm{Cov}(X_i, X_j)\right) \tag{3.63}$$

▪

We showed in this lecture that uncorrelatedness does not imply independence. However, the added condition of joint normality with uncorrelatedness implies independence.

Proof. By Equation 3.63, the joint MGF for random variables $X \sim \textsc{Normal}(\mu_X, \sigma_X^2)$ and $Y \sim \textsc{Normal}(\mu_Y, \sigma_Y^2)$ that are uncorrelated is as follows.

$$\begin{aligned} M_{X,Y}(\mathbf{t}) &= \exp\left(\sum_{i=1}^{2} t_i \mu_i + \sum_{i=1}^{2} \frac{t_i^2}{2} \cdot \mathbb{V}[X_i]\right) && (3.64)\\ &= \exp\left(t_1 \mu_X + t_2 \mu_Y + \frac{t_1^2}{2} \cdot \sigma_X^2 + \frac{t_2^2}{2} \cdot \sigma_Y^2\right) && (3.65)\\ &= \exp\left(t_1 \mu_X + \frac{t_1^2}{2} \cdot \sigma_X^2\right) \cdot \exp\left(t_2 \mu_Y + \frac{t_2^2}{2} \cdot \sigma_Y^2\right) && (3.66)\\ &= M_X(t_1) \cdot M_Y(t_2) && (3.67) \end{aligned}$$

We can conclude that $X \perp\!\!\!\perp Y$ by Property 3.18, since the joint MGF factored into the product of the marginal MGFs. ■

4. Concentration and Tails

In this lecture, we will discuss two most well-known results in probability theory: Law of Large Numbers (LLN) and Central Limit Theorem (CLT). But we will start from two important inequalities that often help to derive bounds on probabilities when only the mean, or both the mean and the variance, of the probability distribution are known.

Agenda:
1. Markov's and Chebyshev's Inequality
2. Law of Large Numbers
3. Monte Carlo Methods
4. Central Limit Theorem

4.1 Markov and Chebyshev's Inequality

We begin with a fundamental **tail** inequality from probability theory. The reason this is referred to as a tail inequality is that gives information about how a random variable behaves away from its mean.

Theorem 4.1 (Markov's Inequality). Let X be a non-negative random variable ($X \geq 0$). Then for any $a > 0$

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Aside: You can convince yourself that equality holds when $X \sim a \cdot \text{Bernouilli}(p)$.

Proof. For continuous X with an associated PDF f_X, we can do the following bound.

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x f_X(x)\, \mathrm{d}x && (4.1)\\
&= \int_0^a x f_X(x)\, \mathrm{d}x + \int_a^\infty x f_X(x)\, \mathrm{d}x && (4.2)\\
&\geq \int_a^\infty x f_X(x)\, \mathrm{d}x \qquad (f_X \geq 0) && (4.3)\\
&\geq \int_a^\infty a f_X(x)\, \mathrm{d}x = a \cdot \mathbb{P}(X \geq a) && (4.4)
\end{aligned}$$

■

If a is small ($a \leq \mathbb{E}[X]$), then the inequality will not be very informative. For $a = \mathbb{E}[X]$, it reduces to

$$\mathbb{P}(X \geq \mathbb{E}[X]) \leq 1. \tag{4.5}$$

However, for large a, we get an informative result that indeed provides useful information. For $a = k \cdot \mathbb{E}[X]$ for $k \gg 1$, it gives

$$\mathbb{P}(X \geq k \cdot \mathbb{E}[X]) \leq k^{-1}. \tag{4.6}$$

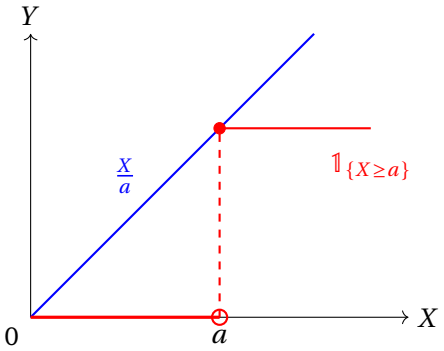
One can also do a visual proof as well, noting that X/a dominates the function $\mathbb{1}_{\{X \geq a\}}$, which yields

$$\mathbb{1}_{\{X \geq a\}} \leq \frac{X}{a} \tag{4.7}$$

By monotonicity of expectation, the inequality is preserved to yields the following.

$$\mathbb{E}[\mathbb{1}_{\{X \geq a\}}] \leq \mathbb{E}\left[\frac{X}{a}\right] \implies \mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \tag{4.8}$$

The essence of this proof is captured in this picture below.



The Chebyshev inequality is a **concentration** inequality, which is because it gives us information on how fast a random variable concentrates to its mean. This inequality in particular is interesting since the Chebyshev inequality is a corollary of the Markov inequality, while Markov was a student for Chebyshev.

Theorem 4.2 (Chebyshev's Inequality). Let X be a random variable with mean μ and variance σ^2. Then for any $a > 0$

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

Aside: Equality here is a little harder to determine, but we have similar intuition from Markov's inequality.

Proof. We directly apply Markov on $Y \equiv (X - \mu)^2$.

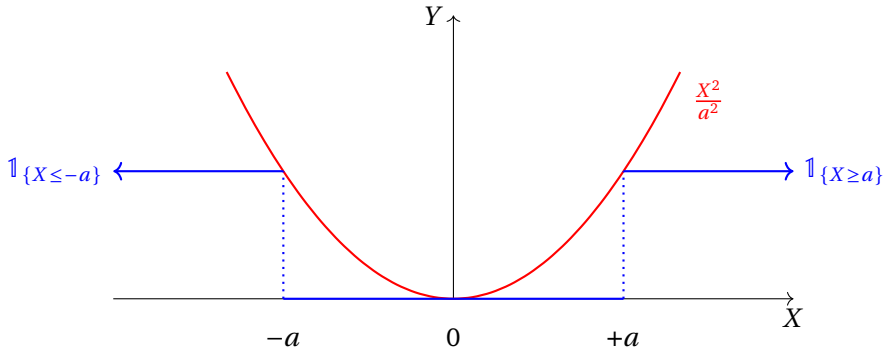
$$\mathbb{P}(|X - \mu| \geq a) = \mathbb{P}(|X - \mu|^2 \geq a^2) \tag{4.9}$$

$$\leq \frac{\mathbb{E}[(X - \mu)^2]}{a^2} \qquad \text{(Markov)} \tag{4.10}$$

$$= \frac{\sigma^2}{a^2} \tag{4.11}$$

■

Once again, we have a summative visual for the inequality below, which can directly show Chebyshev's inequality by monotonicity of expectation.



Example 4.3 Suppose we developed a prediction method, and we want to estimate the unknown error rate p. To this end, we test our prediction method on a set of n new test cases. Let $X_i \sim \text{BERNOULLI}(p)$, where 1 indicates the predictor is wrong, and 0 if right. How does the error rate deviate? ▪

Proof. The error rate would be the sample mean $\overline{X}_n$, defined as

$$\frac{1}{n}\sum_{i=1}^{n} X_i. \tag{4.12}$$

It is natural to use $\overline{X}_n$ as an estimate for p, so we would hope that $\overline{X}_n \approx p$. By linearity of expectation, we have that

$$\mathbb{E}[\overline{X}_n] = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{E}[X_i] = p. \tag{4.13}$$

We can now directly apply Chebyshev's inequality to the sample mean as follows. Intuitively, this tracks the probability that we are off by more than ϵ.

$$\begin{aligned}
\mathbb{P}(|\overline{X}_n - p| \geq \epsilon) &\leq \frac{\mathbb{V}[\overline{X}_n]}{\epsilon^2} && && (4.14)\\
&= \frac{\sum_{i=1}^{n} \mathbb{V}[X_i]}{n^2\epsilon^2} && (X_i \text{ i.i.d}) && (4.15)\\
&= \frac{p(1-p)}{n\epsilon^2} && && (4.16)\\
&\leq \frac{1}{4n\epsilon^2} && (\text{maximized at } p = 1/2) && (4.17)
\end{aligned}$$

This bound can be used for finding n, the number of tests required for having the probability of being off at least by ϵ at most δ. To do this, we can solve the bound $1/(4n\epsilon^2) = \delta$ for n. ■

Another crucial inequality that is used in probability theory is interested in studying how random variables interact with **convex** functions.

Definition 4.4 (Convex Function). A function $f : I \to \mathbb{R}$ on an interval $I \subseteq \mathbb{R}$ is **convex** if

$$f((1-\tau)x + \tau y) \leq (1-\tau)f(x) + \tau f(y)$$

for all $\tau \in [0,1]$ and $x, y \in I$.

Aside: One can write similar results for concave functions.

You can intuitively think of these as functions that lie below its secants and, dually, above its tangents. To obtain some intuition, consider the case when $\tau = 1/2$ for all $x, y \in I$.

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x)+f(y)}{2} \tag{4.18}$$

This says that the function value of the average is at most the average of the function values. This lends itself to a beautiful result known as Jensen's inequality.

Theorem 4.5 (Jensen's Inequality). Let $f : I \to \mathbb{R}$ be a convex function bounded below on an open interval. For every random variable X that takes values in I,

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

4.2 Law of Large Numbers

Now we come to a crowning achievement in probability, which provides a link between randomness and determinism: the law of large numbers.

Theorem 4.6 (Law of Large Numbers (LLN)). Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of i.i.d random variables with finite variance ($\mathbb{V}[X_i] = \sigma^2 < \infty$), and let $\mu = \mathbb{E}[X_i]$. Then, as $n \to \infty$,

$$\overline{X}_n \xrightarrow{\mathbb{P}} \mu$$

Aside: This is formally known as the *weak* law of large numbers.

Intuitively, the LLN says that although the values of the individual X_i are uncertain (random), the value of their average $\overline{X}_n$ becomes more and more certain as n increases.

Before we go any further, we must discuss what the notation $\xrightarrow{\mathbb{P}}$ means. The convergence of a sequence of random variables is far more subtle than in the deterministic setting we say in Ma 1a.

In the deterministic setting, we say that $\{x_n\} \to \mu$ if for all $\epsilon > 0$, there exists an N such that $|x_n - \mu| < \epsilon$ for all $n > N$. For random variables, this formulation makes no sense since $|X_n - \mu| < \epsilon$ is an **event** which occurs with some probability.

As such, we need to update the classical notion of convergence, of which there are several modes (i.e. in probability, distribution, $\mathcal{L}^p$, etc.). We will look at the first one now, and the second later in our discussion of another important theorem.

Definition 4.7 (Convergence in Probability). A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converges **in probability** to μ ($X_n \xrightarrow{\mathbb{P}} \mu$) if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - \mu| \geq \epsilon) = 0$$

Intuitively, $X_n \xrightarrow{\mathbb{P}} \mu$ means that for sufficiently large n, the value of X_n is close to μ with high probability. We now prove the LLN with the tools we have developed thus far.

Proof. We can bound $\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon)$ and take the limit to show the convergence in probability. From previous computations, we know that $\mathbb{E}[\overline{X}_n] = \mu$. Thus, we can directly apply Chebyshev's inequality to show

Aside: Bernoulli's proof was significantly more complicated since he did not know Chebyshev's inequality.

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{\mathbb{V}[\overline{X}_n]}{\epsilon^2} \tag{4.19}$$

$$= \frac{\sum_{i=1}^{n} \mathbb{V}[X_i]}{n^2\epsilon^2} \tag{4.20}$$

$$= \frac{\sigma^2}{n\epsilon^2} \tag{4.21}$$

As $n \to \infty$, the upper bound goes to 0. ∎

It should be noted that the LLN can be proved with much weaker assumptions, such as potentially infinite variance and even non-identical distributions. The beauty of this result is that there are distinctly flavored proofs for each of these settings. This lends credibility to the universality of this beautiful, yet intuitive, result. We will study a quintessential application of the LLN next.

4.3 Monte Carlo Methods

Monte Carlo methods are a wide class of computational algorithms used in all areas of science and engineering. The core idea is to use randomness to solve problems that might be deterministic in principle. This section will cover the most basic application, though some relevant courses that cover other topics include ACM 206 and ACM 112.

Suppose we want to evaluate the integral

$$I = \int_a^b g(x)\,\mathrm{d}x. \tag{4.22}$$

where $g(x)$ is a complicated function such that I admits no analytical representation. The idea is to recast I as an expectation as follows.

$$I = (b-a)\cdot\int_a^b g(x)\cdot\frac{1}{b-a}\,\mathrm{d}x \tag{4.23}$$

$$= (b-a)\cdot\mathbb{E}[g(X)] \qquad (X\sim \textsc{Uniform}[a,b]) \tag{4.24}$$

We can now use the LLN to estimate $\mathbb{E}[g(X)]$ by generating i.i.d samples $\{X_i\}_{i=1}^n \sim \textsc{Uniform}[a,b]$ and averaging. Thus, the LLN says that

$$\frac{1}{n}\sum_{i=1}^n g(X_i) \xrightarrow{\mathbb{P}} \mathbb{E}[g(X)]. \tag{4.25}$$

In total, this lets us conclude that

$$I \approx \frac{b-a}{n}\cdot\sum_{i=1}^n g(X_i). \tag{4.26}$$

One might ask what advantage the Monte Carlo method offers over classical numerical methods (i.e. Simpson's rule, trapezoid rule, etc.) Its power lies in **scalability** to higher dimensions, which most deterministic methods fail due to needing exponentially more data.

$$I = \mathrm{Vol}(V)\cdot\int_V g(\mathbf{x})\cdot\frac{1}{\mathrm{Vol}(V)}\,\mathrm{d}\mathbf{x} \tag{4.27}$$

$$= \mathrm{Vol}(V)\cdot\mathbb{E}[g(\mathbf{X})] \tag{4.28}$$

$$\approx \frac{\mathrm{Vol}(V)}{n}\cdot\sum_{i=1}^n g(\mathbf{X}^i) \qquad (\mathbf{X}^i\sim\textsc{Uniform}(V)) \tag{4.29}$$

A difficulty we have ignored for now is that it is not clear that sampling will be as straightforward in the one-dimensional setting if the domain V is complicated, but there have been modifications to the existing method.

As an immediate extension, there is no reason to restrict our attention to uniform samples. This is known as **importance** sampling since we are weighing samples based on an underlying PDF f.

$$I = \int_a^b \frac{g(x)}{f(x)}\cdot f(x)\,\mathrm{d}x \tag{4.30}$$

$$= \mathbb{E}\left[\frac{g(X)}{f(X)}\right] \tag{4.31}$$

$$\approx \frac{1}{n}\cdot\sum_{i=1}^n \frac{g(X_i)}{f(X_i)} \qquad (X_i\sim f \text{ i.i.d}) \tag{4.32}$$

Aside: Sampling from a generic distribution is a field in its own right!

4.4 Central Limit Theorem

If $X_1, \ldots, X_n$ are random, then the sample mean $\overline{X}_n$ is also random. While the LLN tells us that $\overline{X}_n$ eventually resembles something deterministic for sufficiently large n, it does not tells how $\overline{X}_n$ is distributed. The central limit theorem addresses this. As a note, there are many versions of the CLT with slightly different assumptions and conclusions.

Aside: These objects are referred to as "empirical," since they require observation of samples.

Theorem 4.8 (Lindeberg-Lévy CLT). Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of i.i.d random variables with mean $\mu = \mathbb{E}[X_i]$ and variance $\sigma^2 = \mathbb{V}[X_i]$. Then, as $n \to \infty$,

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

This says that the **standardized** sample mean converges *in distribution* to the standard normal distribution. Once again, we define what this mode of convergence entails, which is compactly stated as pointwise convergence of the CDFs.

Definition 4.9 (Convergence in Distribution). A sequence of random variables $\{Y_n\}_{n=1}^{\infty}$ converges **in distribution** to Y ($Y_n \xrightarrow{\mathcal{D}} Y$) if for all y,

$$\lim_{n\to\infty} F_{Y_n}(y) = F_Y(y)$$

One requires more regularity conditions on the random variables if we would like pointwise convergence of the PDFs or PMFs. Unlike the LLN, the CLT is surprising since **nothing** is assumed about the distribution of $\{X_i\}$ except the existence of the mean and variance. Here are some historical remarks about the progression of the CLT.

1. ($\sim$ 1733) de Moivre stated a CLT for $X_i \sim \text{Bernoulli}(p)$ with $p = 1/2$
2. ($\sim$ 1812) Laplace stated a CLT for $X_i \sim \text{Bernoulli}(p)$ with general p
3. ($\sim$ 1901) Lyapunov rigorously proved the Lindeberg-Lévy CLT.

Aside: Probability was not considered part of math, so the seminal works were not particularly rigorous.

To formulate this directly in terms of probabilities, we have the following where Φ is the standard normal CDF.

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \le x\right) = \Phi(x) \tag{4.33}$$

We now prove the CLT using the powerful objects we discussed last lecture, the MGF!

Proof. Due to the nice interactions of the MGF with sums of independent random variables, we can tractably compute the MGF of $\overline{X}_n$.

$$\begin{aligned}
M_{\overline{X}_n}(t) &= \mathbb{E}[\exp(t\overline{X}_n)] && (4.34)\\
&= \mathbb{E}\left[\exp\left(\frac{t(X_1 + \cdots + X_n)}{n}\right)\right] && (4.35)\\
&= \mathbb{E}\left[\prod_{i=1}^{n} \exp\left(\frac{tX_i}{n}\right)\right] && (4.36)\\
&= \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\frac{tX_i}{n}\right)\right] = \mathbb{E}\left[e^{\frac{tX_1}{n}}\right]^n \qquad (X_i \text{ i.i.d}) && (4.37)
\end{aligned}$$

For sufficiently large n, we can do a second order expansion about μ of the exponential function as follows.

$$M_{\overline{X}_n}(t) \approx \left(\mathbb{E}\left[\exp\left(\frac{t\mu}{n}\right) + \frac{t}{n}\cdot\exp\left(\frac{t\mu}{n}\right)\cdot(X_1-\mu) + \frac{1}{2}\left(\frac{t}{n}\right)^2\cdot\exp\left(\frac{t\mu}{n}\right)\cdot(X_1-\mu)^2\right]\right)^n \tag{4.38}$$

$$= \left(\exp\left(\frac{t\mu}{n}\right) + \frac{1}{2}\left(\frac{t}{n}\right)^2\cdot\exp\left(\frac{t\mu}{n}\right)\cdot\mathbb{V}[X_1]\right)^n \tag{4.39}$$

$$= \exp(t\mu)\cdot\left(1+\frac{t^2\sigma^2}{2n^2}\right)^n \tag{4.40}$$

$$\approx \exp\left(t\mu + \frac{t^2(\sigma/\sqrt{n})^2}{2}\right) \tag{4.41}$$

The last approximation comes from the limit definition of the exponential. This is exactly the MGF of a random variable $X \sim \text{Normal}(\mu, \sigma/\sqrt{n})$. ■

Example 4.10 A tourist in Las Vegas stakes \$1 on each play, with a win paying the customer \$2 and returning their stake with probability $1/4$, and a loss costs the stake with probability $3/4$. What is the probability that the tourist will not lose any money after $n = 100$ games? ▪

Proof. We can represent the total payout as a random variable X defined as

$$X = \sum_{i=1}^{n} X_i \tag{4.42}$$

We can compute the mean and variance as follows.

$$\mathbb{E}[X_i] = \frac{1}{4}\cdot 2 + \frac{3}{4}\cdot(-1) = \frac{-1}{4} \tag{4.43}$$

$$\mathbb{V}[X_i] = \mathbb{E}[X_i^2] - \frac{1}{16} \tag{4.44}$$

$$= \frac{1}{4}\cdot 2^2 + \frac{3}{4}\cdot(-1)^2 - \frac{1}{16} = \frac{27}{16} \tag{4.45}$$

The probability we are interested is $\mathbb{P}(X \geq 0)$, which is equivalent to $\mathbb{P}(\overline{X}_n \geq 0)$. We can now apply the approximation granted by CLT.

$$\mathbb{P}(\overline{X}_n \geq 0) = \mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \geq \frac{\frac{1}{4}}{\frac{\sqrt{27}}{4\cdot\sqrt{100}}}\right) \tag{4.46}$$

$$= \mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \geq \frac{10}{3\sqrt{3}}\right) \tag{4.47}$$

$$= 1 - \mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq \frac{10}{3\sqrt{3}}\right) \tag{4.48}$$

$$\approx 1 - \Phi\left(\frac{10}{3\sqrt{3}}\right) = 0.027 \tag{4.49}$$

■

As a final remark, the LLN and CLT results we presented are **qualitative** in nature, as they do not give a speed of convergence. To obtain quantitative results, one typically needs more moment assumptions on the random variables, and more involved proofs.

5. Conditional Probabilities

One of the most useful concepts in probability theory is **conditioning**, which allows us to use knowledge of events to inform us of the probability of new events.

Agenda:
1. Motivation
2. Discrete Setting
3. Continuous Setting

5.1 Motivation

On a conceptual level, we are often interested in probabilities (or other probabilistic quantities) conditional on certain available information in applications. Informally, we are interested in $\mathbb{P}(A \mid B)$ rather than $\mathbb{P}(A)$. On a technical level, computing certain probabilities is greatly simplified by conditioning on appropriate events or random variables. Informally, $\mathbb{P}(A)$ is difficult, but decomposing using the total law of probability is easier.

We recall Bayes' rule gives us the conditional probability of event A given event B (assuming $\mathbb{P}(B) > 0$), denoted as $\mathbb{P}(A \mid B)$. We now intuitively try to answer what is happening as we move from $\mathbb{P}(A)$ to $\mathbb{P}(A \mid B)$.

Initially, we have the sample space S and probability distribution $\mathbb{P}_S(\cdot) : \{\text{subsets of S}\} \to [0, 1]$. This can be written as follows, with the subscript S referring to the ambient sample space, as shown in **a**).

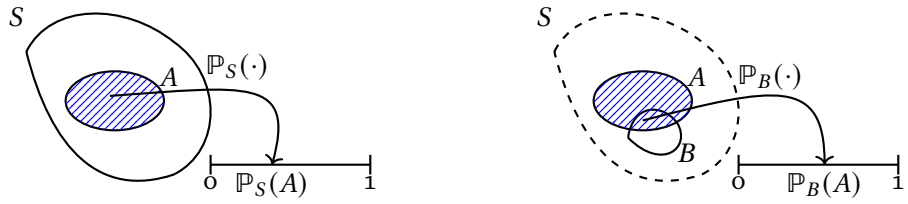
$$\mathbb{P}_S(A) \equiv \mathbb{P}(A \mid S) = \frac{\mathbb{P}(A \cap S)}{\mathbb{P}(S)} = \mathbb{P}(A) \tag{5.1}$$

When we condition on B, we effectively replace the original sample space S by a new sample space $B \subset S$. So, some new information allows us to reduce the original sample space to a smaller one, as shown in **b**).

$$\mathbb{P}_B(A) \equiv \mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{5.2}$$

Aside: This reduction yields a similar relation to that of Equation 5.2.

Thus, $\mathbb{P}_B$ is nothing but a new probability distribution on S. For convenience, we will omit the subscript notation henceforth, simply writing $\mathbb{P}(A \mid B)$.



(a) Full space S

(b) Restricted space $B \subset S$

The picture is useful to keep in mind, and we will now learn how to naturally extend the notion of conditioning to random variables.

5.2 Discrete Setting

We first define the conditional PMF, which will allow us to condition on the realization of a random variable.

Definition 5.1 (Conditional PMF). Let X and Y be discrete random variables. The **conditional PMF** of X given that $Y = y$ is

$$p_{X|Y}(x|y) = \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

Here, $p_{X,Y}$ is the joint PMF of X, Y, and p_Y is the marginal PMF of Y.

The conditional PMF is nothing else but a new (updated) PMF of X in light of the new information that $Y = y$. One can also formulate the equivalent conditional CDF and expectations.

$$F_{X|Y}(x|y) = \sum_{z \le x} p_{X|Y}(z|y) \qquad \mathbb{E}[X \mid Y = y] = \sum_x x \cdot p_{X|Y}(x|y) \tag{5.3}$$

Aside: If $X \perp\!\!\!\perp Y$, we have $p_{X|Y}(x|y) = p_X(x)$ and everything reduces to the unconditional formulae.

Example 5.2 Given the joint PMF $p_{X,Y}(1,1) = 0.5$, $p_{X,Y}(1,2) = 0.1$, $p_{X,Y}(2,1) = 0.1$, $p_{X,Y}(2,2) = 0.3$, find **i**) p_Y, and **ii**) the $p_{X|Y}(x|1)$. ▪

Proof. The marginal of X is computed by summing along a fixed x.

$$p_Y(1) = p_{X,Y}(1,1) + p_{X,Y}(2,1) = 0.6 \tag{5.4}$$
$$p_Y(2) = p_{X,Y}(1,2) + p_{X,Y}(2,2) = 0.4 \tag{5.5}$$

The conditional PMF of X given $Y = 1$ can be computed by the formula in Definition 4.1.

$$p_{X|Y}(1|1) = \frac{p_{X,Y}(1,1)}{p_Y(1)} = \frac{5}{6} \qquad p_{X|Y}(2|1) = \frac{p_{X,Y}(2,1)}{p_Y(1)} = \frac{1}{6} \tag{5.6}$$

■

Example 5.3 Let $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$, where $X \perp\!\!\!\perp Y$. Find $\mathbb{E}[X \mid X + Y = n]$. ▪

Proof. We compute the conditional PMF, noting that $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

$$p_{X|X+Y}(k|n) = \frac{p_{X,X+Y}(k,n)}{p_{X+Y}(n)} \tag{5.7}$$
$$= \frac{p_{X,Y}(k, n-k)}{p_{X+Y}(n)} \tag{5.8}$$
$$= \frac{p_X(k)p_Y(n-k)}{p_{X+Y}(n)} \qquad (X \perp\!\!\!\perp Y) \tag{5.9}$$
$$= \frac{e^{-\lambda_1}\lambda_1^k}{k!} \cdot \frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!} \cdot \frac{n!}{e^{-\lambda_1-\lambda_2} \cdot (\lambda_1+\lambda_2)^n} \tag{5.10}$$
$$= \binom{n}{k} \cdot \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \cdot \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k} \tag{5.11}$$

Therefore, $X \mid X + Y \sim \text{Binomial}(n, p)$, where $p = \lambda_1/(\lambda_1 + \lambda_2)$. Thus,

$$\mathbb{E}[X \mid X + Y = n] = \frac{n\lambda_1}{\lambda_1 + \lambda_2}. \tag{5.12}$$

■

5.3 Continuous Setting

Continuous random variables are defined in terms of PDFs, whose values are not probabilities. Thus, defining a conditional PDF is a bit more trickier. Here is an intuitive motivation, which can be made rigorous using measure theoretic arguments. We recall the probabilistic interpretation of the PDF as

$$\epsilon \cdot f_X(x) \approx \mathbb{P}\left(x - \frac{\epsilon}{2} \le X \le x + \frac{\epsilon}{2}\right). \tag{5.13}$$

We now want to define the conditional PDF of X given that $Y = y$, denoted $f_{X|Y}(x|y)$ for $\epsilon, \delta \ll 1$. We must be careful here since we cannot condition on $Y = y$ as $\mathbb{P}(Y = y) = 0$, so we instead condition on Y being in an infinitesimal interval.

$$\epsilon \cdot f_{X|Y}(x|y) \approx \mathbb{P}\left(x - \frac{\epsilon}{2} \le X \le x + \frac{\epsilon}{2} \,\middle|\, y - \frac{\delta}{2} \le Y \le y + \frac{\delta}{2}\right) \tag{5.14}$$

$$= \frac{\mathbb{P}\left(x - \frac{\epsilon}{2} \le X \le x + \frac{\epsilon}{2}, y - \frac{\delta}{2} \le Y \le y + \frac{\delta}{2}\right)}{\mathbb{P}\left(y - \frac{\delta}{2} \le Y \le y + \frac{\delta}{2}\right)} \tag{5.15}$$

$$\approx \frac{\epsilon\delta \cdot f_{X,Y}(x,y)}{\delta f_Y(y)} \tag{5.16}$$

$$= \epsilon \cdot \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{5.17}$$

This motivates the following definition, with $f_{X|Y}(x|y) = f_X(x)$ if $X \perp\!\!\!\perp Y$.

Definition 5.4 (Conditional PDF). Let X and Y be jointly continuous random variables. The **conditional density** of X given that $Y = y$ is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Here, $f_{X,Y}$ is the joint PDF of X, Y, and f_Y is the marginal PDF of Y.

Analogous to the PMF setting, we can define the conditional CDF and expectations with respect to the PDF.

$$F_{X|Y}(x|y) = \int_{-\infty}^{x} f_{X|Y}(z|y)\,\mathrm{d}z \qquad \mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y) \tag{5.18}$$

Example 5.5 Let $X, Y \sim \textsc{Exponential}(\lambda)$ and $X \perp\!\!\!\perp Y$. Compute $f_{X|X+Y}(x|t)$. ▪

Proof. The joint density is clear since X, Y are independent. However, the conditional density will need an argument via change of variables on the CDF.

$$F_{X,X+Y}(x,t) = \mathbb{P}(X \le x, X + Y \le t) \tag{5.19}$$

$$= \int_{-\infty}^{x}\int_{-\infty}^{t} f_{X,X+Y}(u,v)\,\mathrm{d}v\,\mathrm{d}u \tag{5.20}$$

We can equivalently formulate this directly in terms of the joint PDF of X, Y.

$$F_{X,X+Y}(x,t) = \int_{-\infty}^{x}\int_{-\infty}^{t-u} f_{X,X+Y}(u,v)\,\mathrm{d}v\,\mathrm{d}u \tag{5.21}$$

$$= \int_{-\infty}^{x}\int_{-\infty}^{t} f_{X,Y}(u,w-u)\,\mathrm{d}w\,\mathrm{d}u \qquad (v = w - u) \tag{5.22}$$

Differentiating equations 5.20 and 5.22 twice, we have an equivalence between the two PDFs since the integration bounds match.

$$f_{X,X+Y}(x,t) = f_{X,Y}(x,t-x) \tag{5.23}$$

Using this equivalence, we can use the conditional density formula in DEFINITION 4.4 as follows.

$$f_{X|X+Y}(x|t) = \frac{f_{X,X+Y}(x,t)}{f_{X+Y}(t)} \tag{5.24}$$

$$= \frac{f_{X,Y}(x,t-x)}{f_{X+Y}(t)} \tag{5.25}$$

$$= \frac{(\lambda e^{-\lambda x}) \cdot (\lambda e^{-\lambda(t-x)})}{f_{X+Y}(t)} = \frac{\lambda^2 e^{-\lambda t}}{f_{X+Y}(t)} \tag{5.26}$$

Since the right hand side has no dependence in x (where $x \in [0,t]$), then the conditional distribution is indeed $\text{UNIFORM}[0,t]$, which yields the conditional density as follows.

$$f_{X|X+Y}(x|t) = \frac{1}{t} \tag{5.27}$$

■

By the above, we can compute the density of $f_{X+Y}(t)$ as follows. This is a specific realization of the fact that the sum of n i.i.d exponential random variables with rate λ are distributed according to $\text{GAMMA}(n,\lambda)$.

$$f_{X+Y}(t) = \lambda^2 t e^{-\lambda t} \tag{5.28}$$

Remark 5.6 In general (for independent random variables), one can also compute the density $f_{X+Y}(t)$ through the **convolution** of the two individual densities as shown below.

$$f_{X+Y}(t) = (f_X * f_Y)(t) = \int_{-\infty}^{\infty} f_X(z) \cdot f_Y(t-z)\, \mathrm{d}z \tag{5.29}$$

Aside: This can be extended to the density for a finite sum of **independent** random variables.

Example 5.7 Let $X \sim \text{UNIFORM}[0,1]$, and $Y \mid X = x \sim \text{UNIFORM}[x,1]$. What is the distribution of Y? ▪

Proof. We can compute the joint density $f_{X,Y}(x,y)$ since we are given all the required densities.

$$f_X(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{else} \end{cases} \qquad f_{Y|X}(y|x) = \begin{cases} \frac{1}{1-x} & y \in [x,1] \\ 0 & \text{else} \end{cases} \tag{5.30}$$

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{1-x} & x \in [0,1], y \in [x,1] \\ 0 & \text{else} \end{cases} \tag{5.31}$$

We can now integrate the joint density to obtain the desired marginal.

$$f_Y(y) = \int_0^1 f_{X,Y}(x,y)\, \mathrm{d}x \tag{5.32}$$

$$= \int_0^y \frac{1}{1-x}\, \mathrm{d}x \tag{5.33}$$

$$= -\ln(1-y) \qquad (y \in [0,1]) \tag{5.34}$$

■

6. Conditional Expectations

Last time, we discussed conditional probabilities. In this lecture, we focus on conditional expectations. If X is a random variable, $\mathbb{E}[X]$ is a scalar. If Y is a random variable, $\mathbb{E}[X \mid Y = y]$ is a function of y.

Agenda:
1. Conditional Expectation
2. Analysis of Quicksort
3. Compound Random Variables
4. Conditional Variance

6.1 Conditional Expectation

We define the conditional expectation $\mathbb{E}[X \mid Y]$ in terms of $\mathbb{E}[X \mid Y = y]$.

Definition 6.1 (Conditional Expectation). Let $\mathbb{E}[X \mid Y]$ denote a random variable whose value is $\mathbb{E}[X \mid Y = y]$ when $Y = y$.

If you recall from the first lecture, a random variable Y is a function from the sample space S to $\mathbb{R}$. The conditional expectation is defined for some outcome $\omega \in S$.

$$\mathbb{E}[X \mid Y](\omega) = \mathbb{E}[X \mid Y = Y(\omega)] \tag{6.1}$$

An extremely important property of the conditional expectation is the law of total expectation, which can be intuitively thought of as "cross-averaging."

Theorem 6.2 (Law of Total Expectation).

$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X \mid Y]]$$

The subscript under the expectation here indicates an expectation with respect to a *source of randomness*. In particular, the randomness of $\mathbb{E}[X \mid Y]$ is in X, since Y is conditioned upon (or known). We can write out the total law of expectation explicitly.

Aside: Intuition from linear algebra allows one to interpret the conditional expectation as a projection of a random variable X onto the subspace of information given by Y.

$$\mathbb{E}[X] = \begin{cases} \sum_y \mathbb{E}[X \mid Y = y] \cdot p_Y(y) & (Y \text{ discrete}) \\ \int_{-\infty}^{\infty} \mathbb{E}[X \mid Y = y] \cdot f_Y(y)\, \mathrm{d}y & (Y \text{ continuous}) \end{cases} \tag{6.2}$$

Proof. We can prove the case where X, Y are both continuous.

$$\begin{aligned}
\mathbb{E}_Y[\mathbb{E}_X[X \mid Y]] &= \int_{-\infty}^{\infty} \mathbb{E}_X[X \mid Y = y] \cdot f_Y(y)\, \mathrm{d}y && (6.3)\\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y) \cdot f_Y(y)\, \mathrm{d}x\, \mathrm{d}y && (6.4)\\
&= \int_{-\infty}^{\infty} x \cdot \int_{-\infty}^{\infty} f_{X|Y}(x|y) \cdot f_Y(y)\, \mathrm{d}y\, \mathrm{d}x \qquad \text{(Fubini)} && (6.5)\\
&= \int_{-\infty}^{\infty} x \cdot \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}y\, \mathrm{d}x && (6.6)\\
&= \int_{-\infty}^{\infty} x \cdot f_X(x)\, \mathrm{d}x = \mathbb{E}[X] && (6.7)
\end{aligned}$$

■

We can write this more generally for any function $g(x, y)$.

$$\mathbb{E}_{X,Y}[g(X,Y)] = \mathbb{E}_X[\mathbb{E}_Y[g(X,Y) \mid X]] = \mathbb{E}_Y[\mathbb{E}_X[g(X,Y) \mid Y]] \tag{6.8}$$

To see this exercised in practice, consider an extension of the example from last lecture.

Example 6.3 Let $X \sim \text{Uniform}[0,1]$, $Y \mid X = x \sim \text{Uniform}[x,1]$. What is $\mathbb{E}[Y]$? ▪

Proof. We can do this directly through the conditional expectation.

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[\mathbb{E}[Y \mid X]] && (6.9)\\
&= \mathbb{E}\left[\frac{X+1}{2}\right] && (6.10)\\
&= \frac{1}{4} + \frac{1}{2} = \frac{3}{4} && (6.11)
\end{aligned}$$

This coincides with the computation of computing the expected value normally. ■

In the previous example, conditional expectation simplified the solution, but the problem could still be solved without using conditional expectations (albeit with more computations). Here is an example where conditioning is the only reasonable way to solve the problem.

Example 6.4 A miner is trapped in a mine containing three doors:

- Door 1 leads to a tunnel to safety after t_1 hours of travel.
- Door 2 leads to a tunnel that returns to the mine after t_2 hours of travel.
- Door 3 leads to a tunnel that returns to the mine after t_3 hours of travel.

Find the expected time until the miner reaches safety if each of the doors is chosen uniformly at random. ▪

Proof. If T is the total time spent in the mine, we can use the total law of expectation to condition on the door chosen.

$$\begin{aligned}
\mathbb{E}[T] &= \sum_{i=1}^{3} \mathbb{E}[T \mid D = i] \cdot \mathbb{P}(D = i) && (6.12)\\
&= \frac{1}{3} \cdot (\mathbb{E}[T \mid D = 1] + \mathbb{E}[T \mid D = 2] + \mathbb{E}[T \mid D = 3]) && (6.13)\\
&= \frac{1}{3} \cdot (t_1 + (t_2 + \mathbb{E}[T]) + (t_3 + \mathbb{E}[T])) && (6.14)\\
&= \frac{t_1 + t_2 + t_3}{3} + \frac{2\mathbb{E}[T]}{3} && (6.15)
\end{aligned}$$

Solving for $\mathbb{E}[T]$, we have $\mathbb{E}[T] = t_1 + t_2 + t_3$. ■

The theorem below lists important properties of the conditional expectation.

Theorem 6.5 (Conditional Expectation Properties). Let X, Y, Z be random variables, scalars $a, b \in \mathbb{R}$ and function g.

1. $\mathbb{E}[aX + bY \mid Z] = a \cdot \mathbb{E}[X \mid Z] + b \cdot \mathbb{E}[Y \mid Z]$
2. $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$ if X and Y are independent.
3. $\mathbb{E}[Xg(Y) \mid Y] = g(Y) \cdot \mathbb{E}[X \mid Y]$
4. $\mathbb{E}[X \mid Y, g(Y)] = \mathbb{E}[X \mid Y]$

Aside: The proofs of the results are similar to that of the total law of expectation.

Let's now consider a more serious application of conditioning.

6.2 Analysis of Quicksort

Suppose we want to sort n distinct numbers $\{x_i\}_{i=1}^n$ into the collection $\{x_{(i)}\}_{i=1}^n$, which are the order statistics $(x_{(1)} < \cdots < x_{(n)})$.

Algorithm 6.1 SelectionSort Algorithm

Input: Unordered list $\{x_i\}_{i=1}^n$
for $i = 1, \ldots, n$ **do**
    Scan $n - i$ elements and find i-th smallest $x_{(i)}$

One measure of the efficiency of sorting algorithms is the number of comparisons that the algorithm makes. For the selection-sort algorithm, we must make a large amount of comparisons.

$$M_n = \sum_{i=0}^{n-1} i = \frac{(n-1)n}{2} = \mathcal{O}(n^2) \qquad (n \to \infty) \tag{6.16}$$

There are more efficient algorithms, one of which is the quick sort algorithm.

Algorithm 6.2 QuickSort Algorithm

Input: Unordered list $\{x_i\}_{i=1}^n$
if $n = 2$ **then**
    List $= [x_{(i)}, x_{(j)}]$ ▷ Sort list directly
if $n > 2$ **then**
    Choose $X = x_i$ uniformly at random ▷ Pivot
    Compare X with other $n - 1$ numbers
    Define $S_i = \{x_j \mid x_j < x_i\}, L_i = \{x_j \mid x_j > x_i\}$
    List $= [\textsc{QuickSort}(S_i), x_i, \textsc{QuickSort}(L_i)]$

We can see an example as follows on the unsorted list $\{2, 5, 3, 1, 4\}$.

$$\textsc{QuickSort}(\{2,5,3,1,4\}) = \{\textsc{QuickSort}(\{2,1\}), 3, \textsc{QuickSort}(\{5,4\})\} \tag{6.17}$$
$$= \{1,2,3,4,5\} \tag{6.18}$$

The number of comparisons M_n is a random variable depending on the choice of pivot, which we can do through the alternate formula for the expectation.

$$\mathbb{E}[M_n] = \sum_{j=1}^n \mathbb{E}[M_n \mid X = x_{(j)}] \cdot \mathbb{P}(X = x_{(j)}) \tag{6.19}$$
$$= \sum_{j=1}^n ((n-1) + \mathbb{E}[M_{j-1}] + \mathbb{E}[M_{n-j}]) \cdot \frac{1}{n} \qquad (*) \tag{6.20}$$
$$n\mathbb{E}[M_n] = n \cdot (n-1) + \sum_{j=1}^n (\mathbb{E}[M_{j-1}] + \mathbb{E}[M_{n-j}]) \tag{6.21}$$
$$= n \cdot (n-1) + 2 \cdot \sum_{j=1}^{n-1} \mathbb{E}[M_j] \tag{6.22}$$

The starred step is justified since we do $n - 1$ comparisons on the pivot, and quicksort on the left/right side sets, which carry over the expected value. The last step is justified by symmetry of the summand and $\mathbb{E}[M_0] = 0$.

To compute $\mathbb{E}[M_n]$, we consider a copy of the equation for $n - 1$ as follows.

$$(n-1) \cdot \mathbb{E}[M_{n-1}] = (n-1)(n-2) + 2 \cdot \sum_{j=1}^{n-2} \mathbb{E}[M_j] \tag{6.23}$$

Subtracting Equation 6.23 from Equation 6.22, we have the following.

$$n \cdot \mathbb{E}[M_n] - (n-1) \cdot \mathbb{E}[M_{n-1}] = 2(n-1) + 2 \cdot \mathbb{E}[M_{n-1}] \tag{6.24}$$

$$\mathbb{E}[M_n] = \frac{2(n-1)}{n} + \frac{n+1}{n} \cdot \mathbb{E}[M_{n-1}] \tag{6.25}$$

We can now use this recursively to write an update as follows.

$$\mathbb{E}[M_n] = \frac{2(n-1)}{n} + \frac{n+1}{n} \cdot \mathbb{E}[M_{n-1}] \tag{6.26}$$

$$= \frac{2(n-1)}{n} + \frac{n+1}{n} \cdot \left(\frac{2(n-2)}{n-1} + \frac{n}{n-1} \cdot \mathbb{E}[M_{n-2}] \right) \tag{6.27}$$

$$= \frac{2(n-1)}{n} + \frac{2(n+1)(n-2)}{n(n-1)} + \frac{n+1}{n-1} \cdot \mathbb{E}[M_{n-2}] \tag{6.28}$$

$$= 2 \cdot \sum_{k=1}^{n-1} \frac{(n-k)(n+1)}{(n+1-k)(n+2-k)} \tag{6.29}$$

$$= 2(n+1) \cdot \sum_{i=1}^{n-1} \frac{i}{(i+1)(i+2)} \qquad (i = n-k) \tag{6.30}$$

We can now use partial fraction decomposition and a "continuation" argument to estimate this for sufficiently large n ($n \gg 1$).

Aside: This technique is used extensively by physicists to obtain asymptotics for complicated formulae.

$$\sum_{i=1}^{n-1} \frac{2(n+1) \cdot i}{(i+1)(i+2)} = 2(n+1) \cdot \sum_{i=1}^{n-1} \left(\frac{2}{i+2} - \frac{1}{i+1} \right) \tag{6.31}$$

$$\sim 2(n+1) \cdot \left(\int_3^{n+1} \frac{2}{x} \, \mathrm{d}x - \int_2^n \frac{1}{x} \, \mathrm{d}x \right) \tag{6.32}$$

$$= 2(n+1) \cdot [2\ln(n+1) - 2\ln(3) - \ln(n) + \ln(2)] \tag{6.33}$$

$$\approx 2(n+1) \cdot \left[\ln(n+1) - \ln\left(\frac{n+1}{n} \right) \right] \tag{6.34}$$

$$\approx 2(n+1)\ln(n+1) \tag{6.35}$$

Therefore, we have that the expected number of comparisons is of the complexity $\boxed{\mathcal{O}(n \log n)}$, which is more efficient than the selection sort algorithm.

6.3 Compound Random Variables

A compound random variable is a random variable with "two sources of randomness," in that it is a sum of a random number of random variables. By convention, we define $S = 0$ when $N = 0$.

Definition 6.6 (Compound Random Variable). Let N be a discrete random variable taking integer non-negative values, and $X_1, X_2, \ldots$ be a sequence of (discrete or continuous) random variables. Then $S = \sum_{i=1}^{N} X_i$ is a **compound random variable**.

Compound random variables are used to model many financial, geo-physical, and medical-related phenomena. Think of N as the number of "events" and X_i as the "size" of the i-th event. For example, N is the number of accidents, and X_i are the number of people injured in accident i, which would make S the total number of people injured.

To make the analysis of the compound random variables tractable, the following assumptions are usually made.

1. $X_1, X_2, \ldots$ are i.i.d, as individual event sizes should be independent of each other and all arise from the same distribution
2. $N \perp\!\!\!\perp X_i$ for all i, as the number of events is independent of the size of the constituent events

The expected value of a compound random variable can then be found using the machinery of conditional expectations. Let $\mathbb{E}[N] = \nu$ and $\mathbb{E}[X_i] = \mu$.

$$\begin{aligned}
\mathbb{E}[S] &= \mathbb{E}\left[\sum_{i=1}^{N} X_i\right] && (6.36)\\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{N} X_i \,\middle|\, N\right]\right] && \text{(total law of expectation)} \quad (6.37)\\
&= \mathbb{E}\left[N \cdot \mathbb{E}[X_1 \mid N]\right] && (X_i \text{ i.i.d}) \quad (6.38)\\
&= \mathbb{E}[N \cdot \mu] && (N \perp\!\!\!\perp X_i) \quad (6.39)\\
&= \nu\mu && (6.40)
\end{aligned}$$

A natural question would be if we can find the variance of a compound random variable? We can recast this in terms of the conditional variance, which is defined by analogy with the conditional expectation.

- If X is a random variable, the variance is defined as

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] \tag{6.41}$$

- If Y is another random variable, the conditional variance is defiend as

$$\begin{aligned}
\mathbb{V}[X \mid Y = y] &= \mathbb{E}[(X - \mathbb{E}[X \mid Y = y])^2 \mid Y = y] && (6.42)\\
&= \mathbb{E}[X^2 \mid Y = y] - (\mathbb{E}[X \mid Y = y])^2 && (6.43)
\end{aligned}$$

Definition 6.7 (Conditional Variance). Let $\mathbb{V}[X \mid Y]$ denote a random variable whose value is $\mathbb{V}[X \mid Y = y]$ when $Y = y$.

One can write an analogous total law of variance with the above ideas.

Theorem 6.8 (Law of Total Variance).

$$\mathbb{V}[X] = \mathbb{E}_Y[\mathbb{V}_X[X \mid Y]] + \mathbb{V}_Y[\mathbb{E}_X[X \mid Y]]$$

Aside: There is also a law of total covariance of similar flavor.

Proof. We handle each term separately by utilizing the definitions of the conditional expectation and variance.

$$\begin{aligned}
\mathbb{E}_Y[\mathbb{V}_X[X \mid Y]] &= \mathbb{E}_Y[\mathbb{E}_X[X^2 \mid Y] - (\mathbb{E}_X[X \mid Y])^2] && \text{(conditional variance)}\\
&= \mathbb{E}[X^2] - \mathbb{E}_Y[(\mathbb{E}_X[X \mid Y])^2] && \text{(total law of expectation)}\\
\mathbb{V}_Y[\mathbb{E}_X[X \mid Y]] &= \mathbb{E}_Y[(\mathbb{E}_X[X \mid Y])^2] - (\mathbb{E}_Y[\mathbb{E}_X[X \mid Y]])^2 && \text{(variance)}\\
&= \mathbb{E}_Y[(\mathbb{E}_X[X \mid Y])^2] - \mathbb{E}[X^2] && \text{(total law of expectation)}
\end{aligned}$$

Adding both of these terms yields the desired result. ∎

We can now compute the variance for the compound random variable, further assuming that $\mathbb{V}[N] = \sigma_N^2$ and $\mathbb{V}[X_i] = \sigma_X^2$.

$$\mathbb{V}[S] = \mathbb{E}_N\left[\mathbb{V}_X\left[\sum_{i=1}^N X_i \,\middle|\, N\right]\right] + \mathbb{V}_N\left[\mathbb{E}_X\left[\sum_{i=1}^N X_i \,\middle|\, N\right]\right] \tag{6.44}$$

$$= \mathbb{E}_N\left[N \cdot \mathbb{V}_X[X_1 \mid N]\right] + \mathbb{V}_N\left[N \cdot \mathbb{E}_X[X_1 \mid N]\right] \qquad (X_i \text{ i.i.d}) \tag{6.45}$$

$$= \mathbb{E}_N\left[N \cdot \mathbb{V}_X[X_1]\right] + \mathbb{V}_N\left[N\mu\right] \qquad (N \perp\!\!\!\perp X_i) \tag{6.46}$$

$$= \sigma_X^2 \cdot \mathbb{E}_N[N] + \mu^2 \cdot \mathbb{V}_N[N] \tag{6.47}$$

$$= \sigma_X^2 \cdot \nu + \mu^2 \cdot \sigma_N^2 \tag{6.48}$$

7. More Conditional Probabilities

In the last lecture, we discussed how to compute expectations by conditioning. Here, we will discuss how to compute probabilities by conditioning on appropriate random variables. The link between probabilities and expectations is provided by special random variables, called the **indicator** random variable.

Agenda:
1. Indicator Variables
2. Best Prize Problem
3. Ballot Problem
4. Double Conditioning

7.1 Indicator Random Variables

Let A be any event. The indicator of A is defined as follows.

$$\mathbb{1}_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A \text{ does not occur} \end{cases} \tag{7.1}$$

It is clear that $\mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A)$, and for any random variable X, we have

$$\begin{aligned} \mathbb{E}[\mathbb{1}_A \mid X = x] &= \sum_{a=0}^{1} a \cdot \mathbb{P}_{\mathbb{1}_A \mid X}(a \mid x) && (7.2)\\ &= \mathbb{P}_{\mathbb{1}_A \mid X}(1 \mid x) && (7.3)\\ &= \mathbb{P}(\mathbb{1}_A = 1 \mid X = x) && (7.4)\\ &= \mathbb{P}(A \mid X = x). && (7.5) \end{aligned}$$

These two representations of probabilities in terms of expectations allow us to derive the following useful formula for $\mathbb{P}(A)$.

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{E}[\mathbb{1}_A] && (7.6)\\ &= \mathbb{E}_X[\mathbb{E}[\mathbb{1}_A \mid X]] && (7.7)\\ &= \begin{cases} \sum_x \mathbb{E}[\mathbb{1}_A \mid X = x] \cdot \mathbb{P}(X = x) & \text{if } X \text{ is discrete} \\ \int \mathbb{E}[\mathbb{1}_A \mid X = x] \cdot f_X(x)\, \mathrm{d}x & \text{if } X \text{ is continuous} \end{cases} && (7.8) \end{aligned}$$

We can then obtain the total law of probability by the equivalence established previously.

$$\mathbb{P}(A) = \begin{cases} \sum_x \mathbb{P}(A \mid X = x) \cdot \mathbb{P}(X = x) & \text{if } X \text{ is discrete} \\ \int \mathbb{P}(A \mid X = x) \cdot f_X(x)\, \mathrm{d}x & \text{if } X \text{ is continuous} \end{cases} \tag{7.9}$$

Aside: This should be intuitive, but the above provides the formal justification.

This result allows, in particular, to find $\mathbb{P}(X < Y)$, which often appears in many problems. Let $X \sim f_X$, $Y \sim f_Y$ and $X \perp\!\!\!\perp Y$.

$$\begin{aligned} \mathbb{P}(X < Y) &= \int_{-\infty}^{\infty} \mathbb{P}(X < Y \mid Y = y) \cdot f_Y(y)\, \mathrm{d}y && (7.10)\\ &= \int_{-\infty}^{\infty} \mathbb{P}(X < y) \cdot f_Y(y)\, \mathrm{d}y \qquad (X \perp\!\!\!\perp Y) && (7.11)\\ &= \int_{-\infty}^{\infty} F_X(y) \cdot f_Y(y)\, \mathrm{d}y && (7.12) \end{aligned}$$

Aside: One can obtain a similar result in terms of F_Y, the CDF of Y.

Example 7.1 Let $N \sim \text{Poisson}(\lambda)$ be the number of students who attend ACM 116 lecture. Suppose that each student who attends is independently female w.p p and male w.p $1-p$. Find the joint probability that exactly n female students and m male students attend the lecture today. ▪

Proof. Let N_f be the number of female students, and N_m be the number of male students, with $N = N_f + N_m$. We are interested in computing the following probability, which we can re-express by the total law of probability as

$$\mathbb{P}(N_f = n, N_m = m) = \sum_{k=0}^{\infty} \mathbb{P}(N_f = n, N_m = m \mid N = k) \cdot \mathbb{P}(N = k). \quad (7.13)$$

It is clear that there is only value of k where the probability will be non-zero, which is where $k = n + m$. This simplifies to be

$$\mathbb{P}(N_f = n, N_m = m) = \mathbb{P}(N_f = n, N_m = m \mid N = n + m) \cdot \mathbb{P}(N = n + m). \quad (7.14)$$

Since $N \sim \text{Poisson}(\lambda)$, we can compute the second term in the product as

$$\mathbb{P}(N_f = n, N_m = m) = \mathbb{P}(N_f = n, N_m = m \mid N = n+m) \cdot \frac{\exp(-\lambda) \cdot \lambda^{m+n}}{(m+n)!}. \quad (7.15)$$

Givne $N = n + m$, the two events $\{N_f = n, N_m = m\}$ and $\{N_f = n\}$ are equivalent.

$$\mathbb{P}(N_f = n, N_m = m) = \mathbb{P}(N_f = n \mid N = n + m) \cdot \frac{\exp(-\lambda) \cdot \lambda^{m+n}}{(m+n)!}. \quad (7.16)$$

It is clear that $N_f = n \mid N = n + m \sim \text{Bin}(n + m, p)$, since a student is female with probability p independently of the others. Thus, the final probability simplifies to

$$\mathbb{P}(N_f = n, N_m = m) = \binom{n+m}{n} \cdot p^n \cdot (1-p)^m \cdot \frac{\exp(-\lambda) \cdot \lambda^{m+n}}{(m+n)!}. \quad (7.17)$$

Since the answer is in terms of the problem parameters, we could stop here. However, we massage the expression to obtain insight into the random variables N_f, N_m.

$$\begin{aligned}
\mathbb{P}(N_f = n, N_m = m) &= \frac{(n+m)!}{n!m!} \cdot p^n \cdot (1-p)^m \cdot \frac{\exp(-\lambda) \cdot \lambda^{m+n}}{(m+n)!} && (7.18)\\
&= p^n \cdot (1-p)^m \cdot \frac{\exp(-p\lambda) \cdot \exp(-(1-p)\lambda) \cdot \lambda^{m+n}}{n!m!} && (7.19)\\
&= \underbrace{\frac{(\lambda p)^n \cdot \exp(-p\lambda)}{n!}}_{\sim \text{Poisson}(\lambda p)} \cdot \underbrace{\frac{(\lambda(1-p))^m \cdot \exp(-(1-p)\lambda)}{m!}}_{\sim \text{Poisson}(\lambda(1-p))} && (7.20)
\end{aligned}$$

Since the joint PMF is the product of the marginal PMFs, then we can say $N_f \sim \text{Poisson}(\lambda p)$ and $N_m \sim \text{Poisson}(\lambda(1-p))$, with the key insight being $N_f \perp\!\!\!\perp N_m$. ■

In mathematics, it is useful to think about what problem you actually solved! Here, we established an important theoretical fact.

> **Theorem 7.2 (Poisson Classification).** If each of a Poisson number of events is independently classified either as being type 1 or type 2, the number of type 1 and type 2 events are independent Poisson random variables.

This can be generalized for k-type classification of events.

7.2 Best Prize Problem

Suppose that we are presented with n distinct prizes in a random sequence. Some examples of this may be searching for a gas station to stop for gas, or airline tickets online. After we see prize k, we

- know its relative rank compared to prizes $1, 2, \ldots, k-1$
- must decide whether to accept prize k, or reject it and consider prize $k+1$
- lose the prize if we reject it

The **goal** is to find the optimal stopping rule, or to develop a strategy that maximizes the probability of obtaining the best prize.

Aside: This problem was posed in 1949 and solved in 1958.

It can be shown (non-trivially) through optimal stopping theory that this optimal strategy belongs to the following family of strategies.

Theorem 7.3 (k-strategy). To obtain the best prize in the posed problem,

1. reject the first k prizes
2. accept the first one that is better than all items $\{1, \ldots, k\}$

We can use the law of total probability to compute the probability of selecting the best prize under the k-strategy, which we note as $\mathbb{P}_k(\text{best})$. Here, B represents the true position of the best prize, which each of the n candidates is equally likely to be.

$$\mathbb{P}_k(\text{best}) = \sum_{i=1}^{n} \mathbb{P}_k(\text{best} \mid B = i) \cdot \mathbb{P}(B = i) \tag{7.21}$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{P}_k(\text{best} \mid B = i) \tag{7.22}$$

It is clear that if $i \leq k$, then $\mathbb{P}_k(\text{best} \mid B = i) = 0$ since we reject the first k prizes by the optimal strategy. For $i > k$, the event of selecting the best prize under k-strategy is equivalent tot the event that the best prize among the first $i-1$ prizes is in fact among the first k prizes. Since all of the orderings are equally likely, we have that in this regime, $\mathbb{P}_k(\text{best} \mid B = i) = k/(i-1)$. We can now take a limit of this sum as follows.

$$\mathbb{P}_k(\text{best}) = \frac{1}{n} \cdot \sum_{i=k+1}^{n} \frac{k}{i-1} \tag{7.23}$$

$$\approx \frac{k}{n} \cdot \int_{k}^{n-1} \frac{\mathrm{d}x}{x} \qquad (n \gg 1) \tag{7.24}$$

$$= \frac{k}{n} \cdot \log\left(\frac{n-1}{k}\right) \approx \frac{k}{n} \cdot \log\left(\frac{n}{k}\right) \qquad (n \gg 1) \tag{7.25}$$

By letting $x \approx k/n$ under assumption that $n \gg 1$, the probability function becomes $-x\ln(x)$. We can maximize this by differentiating to yield the following.

$$0 = -\ln(x) - 1 \implies x = 1/e \tag{7.26}$$

Therefore, the optimal value of $k = n/e$, which in turn means that the probability of extracting the optimal candidate is $1/e \approx 37\%$. This is counter-intuitively large, as empirical experiments by psychologists and economists observed that people tended to stop searching too soon.

Aside: This solution is also called the $1/e$-stopping rule.

One drawback of this in applications is that the the value of n must be known *a priori*. However, if n is modeled as a random variable, then a solution exists (albeit, with much more effort). In particular, the optimal probability is much less than $1/e$.

7.3 Ballot Problem

Suppose that in election, candidate A receives n votes, and candidate B receives m votes with $n > m$, i.e. A wins the election. What is the probability that A will always be ahead of B throughout the count of the votes, assuming that all orderings of votes are equally likely?

To see this concretely, consider the case when $n = 3$ and $m = 2$, which yields 10 possible vote counts by elementary combinatorics. The valid counts are highlighted in green, while invalid are highlighted in red.

AAABB AABAB ABAAB BAAAB AABBA

ABABA BAABA ABBAA BABAA BBAAA

Let us denote the event where A remains ahead of B in the count as $\mathrm{A} \gg \mathrm{B}$. We can use induction to show this result as follows, where X denotes the last vote and $\mathbb{P}_{n,m}$ is the probability this event occurs given that A has n votes and B has m.

$$\mathbb{P}_{n,m}(\mathrm{A} \gg \mathrm{B}) = \mathbb{P}_{n,m}(\mathrm{A} \gg \mathrm{B} \mid X = \mathrm{A}) \cdot \mathbb{P}(X = \mathrm{A}) + \mathbb{P}_{n,m}(\mathrm{A} \gg \mathrm{B} \mid X = \mathrm{B}) \cdot \mathbb{P}(X = \mathrm{B}) \tag{7.27}$$

The probability of the last votes are simple since we know the number of votes *a priori*, as well as all orderings being equally likely.

$$\mathbb{P}_{n,m}(\mathrm{A} \gg \mathrm{B}) = \frac{\mathbb{P}_{n,m}(\mathrm{A} \gg \mathrm{B} \mid X = \mathrm{A}) \cdot n}{n+m} + \frac{\mathbb{P}_{n,m}(\mathrm{A} \gg \mathrm{B} \mid X = \mathrm{B}) \cdot m}{n+m} \tag{7.28}$$

However, if the last vote is known, then we can write this recursively as follows.

$$\mathbb{P}_{n,m}(\mathrm{A} \gg \mathrm{B}) = \frac{\mathbb{P}_{n-1,m}(\mathrm{A} \gg \mathrm{B}) \cdot n}{n+m} + \frac{\mathbb{P}_{n,m-1}(\mathrm{A} \gg \mathrm{B}) \cdot m}{n+m} \tag{7.29}$$

We now proceed with our proof by induction by claiming that $\mathbb{P}(\mathrm{A} \gg \mathrm{B}) = (n - m)/(n + m)$. This is trivially satisfied in the case where $n = 1$ and $m = 0$. By the induction hypothesis, if we assume this holds for all n, m such that $n + m = k$, then we can use the above recursive relation to prove the result.

$$\mathbb{P}_{n,m}(\mathrm{A} \gg \mathrm{B}) = \frac{n}{n+m} \cdot \frac{(n-1)-m}{(n-1)+m} + \frac{m}{n+m} \cdot \frac{n-(m-1)}{n+(m-1)} \tag{7.30}$$

$$= \frac{n^2 - n - nm + mn - m^2 + m}{(n+m)(n+m-1)} \tag{7.31}$$

$$= \frac{n^2 - n - m^2 + m}{(n+m)(n+m-1)} \tag{7.32}$$

$$= \frac{(n-m)(n+m-1)}{(n+m)(n+m-1)} \tag{7.33}$$

$$= \frac{n-m}{n+m} \tag{7.34}$$

In the case that ties are allowed, this probability becomes the following.

$$\mathbb{P}_{m,n}(\mathrm{A} \gg \mathrm{B}) = \frac{n-m+1}{n+m} \tag{7.35}$$

There are many variants of this problem, each of which admit many solutions. Conditioning is a ubiquitous concept in probability, and this example shows the basic framework of these arguments.

7.4 Double Conditioning

As we know, computing expectations $\mathbb{E}[X]$ can be greatly simplified by conditioning on an appropriate random variable, i.e. $\mathbb{E}[X]$ can be computed in terms of $\mathbb{E}[X \mid Y = y]$. But $\mathbb{E}[X \mid Y = y]$ is just an ordinary expectation that is just computed with respect to $p_{X|Y}$ or $f_{X|Y}$.

As such, conditional expectations satisfy all the properties of ordinary expectations. In particular, we can simplify the computation of $\mathbb{E}[X \mid Y = y]$ by conditioning on an "intermediate" random variable Z.

$$\mathbb{E}[X \mid Y = y] = \begin{cases} \sum_z \mathbb{E}[X \mid Y = y, Z = z] \cdot \mathbb{P}(Z = z \mid Y = y) \\ \int \mathbb{E}[X \mid Y = y, Z = z] \cdot f_{Z|Y}(z \mid y)\, \mathrm{d}z \end{cases} \tag{7.36}$$

Aside: Compare this to the law of total probability and expectation in the previous lecture.

Recast more compactly, we can write this as follows. Note that subscript of the expectation denotes what randomness it is taken with respect to.

$$\mathbb{E}_X[X \mid Y] = \mathbb{E}_Z[\mathbb{E}_X[X \mid Y, Z] \mid Y] \tag{7.37}$$

We also get the double conditioning principle for free on probabilities due to the connection given by indicator random variables.

$$\mathbb{P}(A \mid X = x) = \begin{cases} \sum_y \mathbb{P}(A \mid X = x, Y = y) \cdot \mathbb{P}(Y = y \mid X = x) \\ \int \mathbb{P}(A \mid X = x, Y = y) \cdot f_{Y|X}(y \mid x)\, \mathrm{d}y \end{cases} \tag{7.38}$$

In short, we can write the latter as follows.

$$\mathbb{P}(A \mid X) = \mathbb{E}_Y[\mathbb{P}(A \mid X, Y) \mid X] \tag{7.39}$$

Example 7.4 A car insurance company classifies each of its policy holders as being one of k risk types : $1, \ldots, k$. Let N_i be the number of accidents that a type i policy holder has in a year, $N_i \sim \text{Poisson}(\lambda_i)$. It is estimated from the previous data that a new customer has risk type i with probability p_i. A new customer had n accidents in this first year. Compute the **1**) expected number of accidents in the second year, and **2**) probability that he will have m accidents in his second year. ▪

Proof. For **1**), we compute the conditional expectation using the double conditioning principle as follows, where $N^{(i)}$ represents the number of accidents in year i and T is the type.

$$\mathbb{E}[N^{(2)} \mid N^{(1)} = n] = \sum_{i=1}^{k} \mathbb{E}[N^{(2)} \mid N^{(1)} = n, T = i] \cdot \mathbb{P}(T = i \mid N^{(1)} = n) \tag{7.40}$$

We can compute the conditional probability via Bayes' formula, noting that $N^{(1)} = n \mid T = i \sim \text{Poisson}(\lambda_i)$.

$$\mathbb{P}(T = i \mid N^{(1)} = n) = \frac{\mathbb{P}(N^{(1)} = n \mid T = i) \cdot \mathbb{P}(T = i)}{\mathbb{P}(N^{(1)} = n)} \tag{7.41}$$

$$= \frac{\mathbb{P}(N^{(1)} = n \mid T = i) \cdot \mathbb{P}(T = i)}{\sum_{j=1}^{k} \mathbb{P}(N^{(1)} = n \mid T = j) \cdot \mathbb{P}(T = j)} \tag{7.42}$$

$$= \frac{\exp(-\lambda_i) \cdot \lambda_i^n \cdot p_i}{\sum_{j=1}^{k} \exp(-\lambda_j) \cdot \lambda_j^n \cdot p_j} \tag{7.43}$$

Returning the original expression, we know that conditioned on $T = i$, $N^{(2)} \perp\!\!\!\perp N^{(1)}$, and in particular, $N^{(2)} \sim \text{Poisson}(\lambda_i)$. Therefore, the expectation is simply λ_i, which yields the following.

$$\mathbb{E}[N^{(2)} \mid N^{(1)} = n] = \sum_{i=1}^{k} \lambda_i \cdot \frac{\exp(-\lambda_i) \cdot \lambda_i^n \cdot p_i}{\sum_{j=1}^{k} \exp(-\lambda_j) \cdot \lambda_j^n \cdot p_j} \tag{7.44}$$

$$= \frac{\sum_{i=1}^{k} \exp(-\lambda_i) \cdot \lambda_i^{n+1} \cdot p_i}{\sum_{j=1}^{k} \exp(-\lambda_j) \cdot \lambda_j^n \cdot p_j} \tag{7.45}$$

For **2**), we can do something nearly identical through the double conditioning principle as follows.

$$\mathbb{P}[N^{(2)} = m \,|\, N^{(1)} = n] = \sum_{i=1}^{k} \mathbb{P}[N^{(2)} = m \,|\, N^{(1)} = n, T = i] \cdot \mathbb{P}(T = i \,|\, N^{(1)} = n) \tag{7.46}$$

The second probability does not change, and we can again use the fact that given the type $T = i$, the number of accidents that a type i customer has in successive years are independent Poisson random variables.

$$\mathbb{P}[N^{(2)} = m \mid N^{(1)} = n] = \sum_{i=1}^{k} \frac{\exp(-\lambda_i) \cdot \lambda_i^m}{m!} \cdot \frac{\exp(-\lambda_i) \cdot \lambda_i^n \cdot p_i}{\sum_{j=1}^{k} \exp(-\lambda_j) \cdot \lambda_j^n \cdot p_j} \tag{7.47}$$

$$= \frac{\sum_{i=1}^{k} \exp(-2\lambda_i) \cdot \lambda_i^{n+m} \cdot p_i}{m! \cdot \sum_{j=1}^{k} \exp(-\lambda_j) \cdot \lambda_j^n \cdot p_j} \tag{7.48}$$

■

8. Random Vectors

In practice, the parameters we are interested in discovering the distribution of are high-dimensional random quantities. Thus, we will take this lecture to extend the notions of distributions and operators using the support of linear algebra in order to do matrix computations.

Agenda:
1. Review of Matrix Algebra
2. Random Vectors and Matrices
3. Covariance Matrix
4. Karhunen-Loève Expansion
5. Transformations of Random Vectors

8.1 Review of Matrix Algebra

First, we will recall the definition of a matrix as follows.

$$A = \begin{bmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ a_{i1} & \cdots & \boxed{a_{ij}} & \cdots & a_{in} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mj} & \dots & a_{mn} \end{bmatrix}$$

Any element in A can be identified by a_{ij}, in which i corresponds to the i-th row of A, and j to the j-th column of A. This is also sometimes referred to as the (i, j)-entry of A.

Naturally, we want to see if our traditional operations on numbers work in the context of matrices. For the most part, yes!

Property 8.1 (Matrix Operations). For matrices $A, B, C \in \mathbb{M}_{m\times n}$, we can define addition.

$$C = A + B \implies c_{ij} = a_{ij} + b_{ij} \tag{8.1}$$

Addition of matrices is both commutative and associative.

$$\begin{aligned} A + B &= B + A \\ (A + B) + C &= A + (B + C) \end{aligned} \tag{8.2}$$

We can also multiply matrices by a scalar $\alpha \in \mathbb{R}$.

$$B = \alpha A \implies b_{ij} = \alpha \cdot a_{ij} \tag{8.3}$$

Finally, we can multiply matrices together to form a new matrix, for $A \in \mathbb{M}_{m\times n}$, $B \in \mathbb{M}_{n\times p}$, $C \in \mathbb{M}_{m\times p}$. In English, the (i, j)-entry of C is simply the dot product of the i-th row of A and the j-th column of B.

$$C = AB \implies c_{ij} = \sum_{k=1}^{n} a_{ik} \cdot b_{kj} \tag{8.4}$$

In general, a lot of the natural properties we expect to hold matrices do, such as associativity and distributivity of multiplication and addition. However, there is one **critical** property that is not preserved when talking about matrices.

Aside: One may ask why we do not define multiplication of matrices on an element level as follows.

$$c_{ij} = a_{ij} \cdot b_{ij}$$

While this is properly defined (Hadamard product), this is quite useless in most of the problems we will look at.

Warning 8.2 With respect to matrix multiplication, we do not generally have the property of commutativity.

$$AB \neq BA$$

As such, a lot of the identities we will discuss involving products of matrices must be treated with care! ▪

We can introduce a fundamentally new operation on matrices called the **transposition**, or an interchanging of the rows and columns.

Definition 8.3 (Transposition). If $A \in \mathbb{M}_{m\times n}$, then the **transpose** of this matrix is denoted by $A^\top \in \mathbb{M}_{n\times m}$ and is defined element-wise as follows.

$$(A^\top)_{ij} = A_{ji}$$

Aside: In the case of a square matrices, $A^\top$ can be visualized as "reflecting" the matrix entries across the main diagonal. Otherwise, transposition can be thought of turning flat matrices ($n > m$) into tall matrices ($n < m$) and vice versa.

In summary, the rows of A are the columns of $A^\top$ and columns of A are the rows of $A^\top$. We note the following basic properties about the transpose of a matrix as follows.

Property 8.4 (Transpose of Transpose).

$$(A^\top)^\top = A$$

Property 8.5 (Transpose of Sum).

$$(A+B)^\top = A^\top + B^\top$$

Property 8.6 (Transpose of Scalar Multiple).

$$(\alpha A)^\top = \alpha A^\top$$

Property 8.7 (Transpose of Product).

$$(A_1 \cdots A_k)^\top = A_k^\top \cdots A_1^\top$$

We can extend multiplication to vectors through the inner product.

Definition 8.8 (Euclidean Inner Product). If $x, y \in \mathbb{R}^n$, then the **dot** product of x, y is as follows.

$$\langle x, y\rangle = x^\top y = \sum_{i=1}^n x_i y_i$$

The inner product can help us define the length of a vector through the norm.

Definition 8.9 (Euclidean Norm). If $x \in \mathbb{R}^n$, then the **norm** of x is defined as follows:

$$\|x\| = \sqrt{\langle x, x\rangle} = \sqrt{\sum_{i=1}^n x_i^2}$$

Aside: More generally, one should note that every inner product induces a norm via the first equality, but not every norm can come from an inner product.

Finally, we can relate the two ideas via the Cauchy-Schwarz inequality.

Theorem 8.10 (Cauchy-Schwarz Inequality). For any vectors $x, y \in \mathbb{R}^n$, the following is true, with equality iff $y = \alpha x$.

$$|\langle x, y\rangle| \leq \|x\| \cdot \|y\|$$

8.2 Random Vectors and Matrices

With the preface of linear algebra completed, we can proceed to define their random analogs.

Definition 8.11 (Random Vectors and Matrices). A vector (resp. matrix) whose entries are random variables is a **random vector** (resp. matrix).

Aside: This can also be generalized to randomized tensors, which are of particular importance in matrix methods.

Many of the concepts introduced for random variables can be naturally extended to these random vectors/matrices.

Definition 8.12 For a random vector $\mathbf{X} = (X_1, \cdots, X_n)^\top$, $\mathbb{E}[\mathbf{X}]$ is defined as follows.

$$\mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], \cdots, \mathbb{E}[X_n])^\top$$

For a random matrix $\mathbf{X}$, it is defined as follows.

$$(\mathbb{E}[\mathbf{X}])_{ij} = \mathbb{E}[\mathbf{X}_{ij}]$$

We will consider an example below in which we have a random and deterministic matrix interact.

Property 8.13 Let $\mathbf{X} = (X_1, \cdots, X_n)^\top$ be a random vector, and $A \in \mathbb{M}_{m\times n}$ that is deterministic. Show that the following is true.

$$\mathbb{E}[A\mathbf{X}] = A \cdot \mathbb{E}[\mathbf{X}] \tag{8.5}$$

Proof. We can show this equality component-wise by matrix multiplication as follows.

$$\begin{aligned}
(\mathbb{E}[A\mathbf{X}])_i &= \mathbb{E}\left[\sum_{j=1}^n a_{ij} \cdot \mathbf{X}_j\right] && (8.6)\\
&= \sum_{j=1}^n a_{ij} \cdot \mathbb{E}\left[\mathbf{X}_j\right] && (8.7)\\
&= \sum_{j=1}^n a_{ij} \cdot \mathbb{E}\left[\mathbf{X}\right]_j && (8.8)\\
&= (A \cdot \mathbb{E}[\mathbf{X}])_i && (8.9)
\end{aligned}$$

■

Property 8.14 If A, B, C are deterministic and $\mathbf{X}$ is a random matrix, then

$$\mathbb{E}[A\mathbf{X}B + C] = A\mathbb{E}[\mathbf{X}]B + C \tag{8.10}$$

If one wanted to generalize the variance operator for random vectors/matrices, a little more nuance is required, which we will discuss in the next section.

Example 8.15 (Trace Estimation). In applied mathematics, sometimes we have large matrices $A \in \mathbb{M}_{n\times n}$ for which we cannot access the full contents of the matrix, but instead only have a function that gives $u \mapsto Au$. However, we are interested in global quantities about the matrix A, such as the trace of A.

Aside: Professor Tropp teaches a wonderful graduate class on randomized algorithms for matrix computations, which covers this algorithm and many more of these methods.

$$\operatorname{tr}(A) = \sum_{i=1}^n a_{ii} \tag{8.11}$$

One can do this by doing a Monte Carlo simulation of the quantity x below with $\mathbf{Z} = (Z_1, \cdots, Z_n)$ where $Z_i \sim$ Rademacher i.i.d.

$$x = \mathbf{Z}^\top A\mathbf{Z} \implies \mathbb{E}[x] = \operatorname{tr}(A) \tag{8.12}$$

▪

8.3 Covariance Matrix

Definition 8.16 Let $\mathbf{X} = (X_1, \cdots, X_n)^\top$ be a random vector with mean vector $\mu = \mathbb{E}[\mathbf{X}]$. We then define the **covariance** matrix of $\mathbf{X}$ as follows.

$$\Sigma_{\mathbf{X}} = \text{Cov}(\mathbf{X}, \mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mu) \cdot (\mathbf{X} - \mu)^\top]$$

Aside: This can be interpreted as the *outer* product of the vector $\mathbf{X}$ with itself, which is typically denoted as $\mathbf{X} \otimes \mathbf{X} = \mathbf{X} \cdot \mathbf{X}^\top$.

By the definition of matrix multiplication, one can verify that this is indeed a matrix which looks like the following upon inspection.

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} \mathbb{V}[X_1] & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \mathbb{V}[X_2] & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \cdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \mathbb{V}[X_n] \end{bmatrix}$$

One can view this as the generalization of the variance $\mathbb{V}[X]$ for higher dimensions, though most people in probability theory typically refer to it as the covariance matrix due to fact that $\mathbb{V}[X] = \text{Cov}(X, X)$. We will now discuss some properties of the covariance matrix as follows.

Property 8.17 The covariance matrix $\Sigma_{\mathbf{X}}$ is symmetric.

Property 8.18 If $(\Sigma_{\mathbf{X}})_{ij} = 0$, then $X_i \perp X_j$ (they are uncorrelated).

Property 8.19 The covariance matrix can be written as follows.

$$\Sigma_{\mathbf{X}} = \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \mu\mu^\top \tag{8.13}$$

Proof. Expanding the insides of the expectation in the definition and utilizing that $\mathbf{X}$ is the only random quantity such that $\mathbb{E}[\mathbf{X}] = \mu$, we have the following.

$$\begin{aligned} \Sigma_{\mathbf{X}} &= \mathbb{E}[\mathbf{X}\mathbf{X}^\top - \mathbf{X}\mu^\top - \mu\mathbf{X}^\top + \mu\mu^\top] && (8.14)\\ &= \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \mathbb{E}[\mathbf{X}\mu^\top] - \mathbb{E}[\mu\mathbf{X}^\top] + \mathbb{E}[\mu\mu^\top] && (8.15)\\ &= \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \mu\mu^\top - \mu\mu^\top + \mu\mu^\top && (8.16)\\ &= \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \mu\mu^\top && (8.17) \end{aligned}$$

■

Property 8.20 If $\mathbf{X}$ is a random vector and A is a deterministic matrix, then the covariance matrix of $\mathbf{Y} = A\mathbf{X}$ is

$$\Sigma_{\mathbf{Y}} = A\Sigma_{\mathbf{X}}A^\top$$

Proof. We can compute this through the covariance formula and substituting the definition of $\mathbf{Y} = A\mathbf{X}$ and $\mathbb{E}[\mathbf{Y}] = A\mu$ as follows.

$$\begin{aligned} \Sigma_{\mathbf{Y}} &= \mathbb{E}[(\mathbf{Y} - \mu_{\mathbf{Y}}) \cdot (\mathbf{Y} - \mu_{\mathbf{Y}})^\top] && (8.18)\\ &= \mathbb{E}[(A\mathbf{X} - A\mu) \cdot (A\mathbf{X} - A\mu)^\top] && (8.19)\\ &= \mathbb{E}[A(\mathbf{X} - \mu) \cdot (\mathbf{X} - \mu)^\top A^\top] && (8.20)\\ &= A\mathbb{E}[(\mathbf{X} - \mu) \cdot (\mathbf{X} - \mu)^\top]A^\top && (8.21)\\ &= A\Sigma_{\mathbf{X}}A^\top && (8.22) \end{aligned}$$

■

Property 8.21 $\Sigma_{\mathbf{X}} \succcurlyeq 0$, or $\Sigma_{\mathbf{X}}$ is positive semi-definite.

Proof. Consider a random variable $\mathbf{Y} = a^\top \mathbf{X}$, where $a \in \mathbb{R}^n$ is arbitrary. Then, using Property 8.20, we obtain the desired inequality.

$$0 \leq \mathbb{V}[Y] = \Sigma_Y = a^T \Sigma_X a \tag{8.23}$$

■

Aside: Recall that a matrix M is called positive semi-definite if 1) M is symmetric and 2) for all $a \in \mathbb{R}^n$ the quadratic from

$$a^\top M a \geq 0$$

Property 8.22 If Σ_X is singular, then one of $\{X_i\}$ is a linear function of the other components.

Proof. Since $\Sigma_{\mathbf{X}}$ is singular, there exists a non-zero vector $a \in \mathbb{R}^n$ such that $\Sigma_{\mathbf{X}} a = 0$. Consider a random variable $\mathbf{Y} = a^T \mathbf{X}$, then its variance (again using Property 9.20) is

$$\mathbb{V}[\mathbf{Y}] = \Sigma_{\mathbf{Y}} = a^T \Sigma_{\mathbf{X}} a = 0 \tag{8.24}$$

Thus, $\mathbf{Y}$ is a constant. Let $\mathbf{Y} = \mu_{\mathbf{Y}}$, which is constant. Then

$$\mu_Y = a_1 X_1 + \ldots + a_n X_n \tag{8.25}$$

Since a is non-zero, one of the components of X is a linear function of the remaining components. ■

Aside: This property once again lends itself nicely to the interpretation of a vector space of random variables.

If we are instead interested in comparing two *different* random vectors, we can generalize the covariance matrix definition as follows.

Definition 8.23 If $\mathbf{X} = (X_1, \cdots, X_n)^\top$ and $\mathbf{Y} = (Y_1, \cdots, Y_m)^\top$ are two random vectors, then the **cross-covariance** matrix is defined by the following.

$$\Sigma_{\mathbf{X},\mathbf{Y}} = \mathrm{Cov}(X, Y) = \mathbb{E}[(\mathbf{X} - \mu_{\mathbf{X}}) \cdot (\mathbf{Y} - \mu_{\mathbf{Y}})^\top] \in \mathbb{M}_{n \times m}$$

One can equivalently define the full covariance matrix for the "stacked" vector $\mathbf{Z} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}$ as follows. Note that $\Sigma_{\mathbf{Z}}$ is a block matrix that is in the space $\mathbb{M}_{(n+m)\times(n+m)}$.

$$\Sigma_{\mathbf{Z}} = \begin{bmatrix} \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{X},\mathbf{Y}} \\ \Sigma_{\mathbf{Y},\mathbf{X}} & \Sigma_{\mathbf{Y}} \end{bmatrix} \tag{8.26}$$

There are also more generic properties for cross-covariance matrices.

Property 8.24

$$\Sigma_{\mathbf{X},\mathbf{Y}} = (\Sigma_{\mathbf{Y},\mathbf{X}})^\top$$

Property 8.25

$$\Sigma_{\mathbf{X},\mathbf{Y}_1+\mathbf{Y}_2} = \Sigma_{\mathbf{X},\mathbf{Y}_1} + \Sigma_{\mathbf{X},\mathbf{Y}_2}$$

Property 8.26

$$\Sigma_{\mathbf{X}+\mathbf{Y}} = \Sigma_{\mathbf{X}} + \Sigma_{\mathbf{X},\mathbf{Y}} + \Sigma_{\mathbf{Y},\mathbf{X}} + \Sigma_{\mathbf{Y}}$$

8.4 Karhunen-Loève Expansion

Consider a random vector $\mathbf{X}$ with associated covariance matrix $\Sigma_{\mathbf{X}}$. It is clear that if the components of $\mathbf{X}$ are pairwise uncorrelated, then the covariance matrix will be diagonal due to the following structure.

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} \mathbb{V}[X_1] & 0 & \cdots & 0 \\ 0 & \mathbb{V}[X_2] & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbb{V}[X_n] \end{bmatrix} \tag{8.27}$$

This is doubly useful since diagonal matrices are very easy to work with, as well as uncorrelated components being useful in practice. Thus, it would be useful if there was a transformation of a random vector that yielded another random vector whose components are uncorrelated.

Theorem 8.27 (Karhunen-Loève Expansion). For any random vector $\mathbf{X}$, there exists an *orthogonal* matrix Q such that $\mathbf{Y}$ has uncorrelated components ($Y_i \perp Y_j$).

$$\mathbf{X} = Q\mathbf{Y}$$

Aside: There is an infinite-dimensional generalization of this theorem for stochastic processes, but this requires some knowledge of functional analysis to understand.

Recall that an orthogonal matrix is such that $Q^\top Q = QQ^\top = I$, or equivalently that $Q^{-1} = Q^\top$. One can see this expansion write above looks like the following element-wise for all $i \in [n]$ if $x = (X_1, \cdots, X_n)^\top$.

$$X_i = \sum_{j=1}^{n} q_{ij} \cdot Y_j \tag{8.28}$$

Aside: To those who have taken ACM 95ab, do you see a resemblance to a Fourier decomposition?

We will now prove this result using the **spectral theorem**.

Proof. Since we proved $\Sigma_{\mathbf{X}}$ is symmetric, then the spectral theorem says that this matrix admits an orthogonal decomposition as follows.

$$\Sigma_{\mathbf{X}} = Q\Lambda Q^\top \tag{8.29}$$

where $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_n)$ is a diagonal matrix with the eigenvalues of the covariance matrix, and Q is a matrix whose column vectors are the corresponding eigenvectors to the eigenvalues.

Furthermore, we showed in Property 8.20 that the covariance matrices of $\mathbf{Y}$ and $\mathbf{X} = Q\mathbf{Y}$ are related as follows.

$$\Sigma_{\mathbf{X}} = Q\Sigma_{\mathbf{Y}}Q^\top \tag{8.30}$$

Thus, we can isolate for $\Sigma_{\mathbf{Y}}$ as follows.

$$\Sigma_{\mathbf{Y}} = Q^\top \Sigma_{\mathbf{X}} Q = (Q^\top Q)\Lambda(Q^\top Q) = \Lambda \tag{8.31}$$

Thus, the covariance matrix of $\mathbf{Y}$ is exactly a diagonal matrix of the eigenvalues, thusly implying that $\mathbf{Y}$ has uncorrelated components. ■

Oftentimes, $Q^\top$ is called the *decorrelating* transformation, since one can recover the uncorrelated $\mathbf{Y}$ from the correlated $\mathbf{X}$.

$$\mathbf{Y} = Q^\top \mathbf{X} \tag{8.32}$$

8.5 Transformations of Random Vectors

Let $\mathbf{X} = (X_1, \cdots, X_n)$ be a random vector, and let $G : \mathbb{R}^n \to \mathbb{R}^n$ be a vector-valued function. Recall from multi-variable calculus that a vector-valued function takes on the following form component-wise.

$$y_1 = g_1(x_1, \cdots, x_n) \quad \cdots \quad y_n = g_n(x_1, \cdots, x_n) \tag{8.33}$$

Since we know $\mathbf{X}$, we can easily write down its joint distribution. If we instead consider the quantity $\mathbf{Y} = G(\mathbf{X})$, can we determine what the joint distribution of $\mathbf{Y}$ is?

The short answer is that we **can** determine the joint distribution of $\mathbf{Y}$, under some mild regularity conditions of G, namely that it is invertible and differentiable. Furthermore, for the aforementioned well-behaved G, we can analytically show the joint density.

Aside: Formally, we need that G is a $\mathfrak{C}^1$-diffeomorphism, which means it is continuously differentiable and has a continuous inverse.

We can prove this by using the change-of-variables approach from multi-variable calculus on the densities. Consider the probability we are interested in writing down below, and how we can interpret this probability in the context of the known random vector $\mathbf{X}$ as follows.

$$\int_{\mathsf{A}} f_{\mathbf{Y}}(y)\, \mathrm{d}y = \mathbb{P}(\mathbf{Y} \in \mathsf{A}) = \mathbb{P}(\mathbf{X} \in G^{-1}(\mathsf{A})) = \int_{G^{-1}(\mathsf{A})} f_{\mathbf{X}}(x)\, \mathrm{d}x \tag{8.34}$$

Our integrand is now known, but we unfortunately do not know what the pre-image of A is. This then motivates a change of variable on the integral with respect to $x = G^{-1}(y)$. For brevity's sake, we will say $H = G^{-1}$, which implies $x = H(y)$.

Aside: Note that $G^{-1}(\mathsf{A})$ is the pre-image of the set A, which consists of all points in $\mathbb{R}^n$ such that mapping through G yields a point in A.

$$x_1 = h_1(y_1, \cdots, y_n) \quad \cdots \quad x_n = h_n(y_1, \cdots, y_n) \tag{8.35}$$

This yields the following integral upon changing the bounds.

$$\mathbb{P}(\mathbf{Y} \in \mathsf{A}) = \int_{\mathsf{A}} f_{\mathbf{X}}(H(y))\, \mathrm{d}x \tag{8.36}$$

We now need to handle the differential term, which can be done by recalling the following formula in terms of the Jacobian of the mapping H.

$$\mathrm{d}x = |\det J_H(y)|\, \mathrm{d}y \tag{8.37}$$

Recall the definition of the Jacobian below for $H : \mathbb{R}^n \to \mathbb{R}^n$, in which the component functions are defined as above.

$$J_H(y) = \begin{bmatrix} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_1}{\partial y_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial h_n}{\partial y_1} & \cdots & \frac{\partial h_n}{\partial y_n} \end{bmatrix} \tag{8.38}$$

Thus, our integral becomes the following, which we can equate to our initial formulation.

$$\int_{\mathsf{A}} f_{\mathbf{Y}}(y)\, \mathrm{d}y = \mathbb{P}(\mathbf{Y} \in \mathsf{A}) = \int_{\mathsf{A}} f_{\mathbf{X}}(H(y))\, |\det J_H(y)|\, \mathrm{d}y \tag{8.39}$$

We can then equate the integrands to then write down the formula for the joint distribution of $\mathbf{Y}$.

Theorem 8.28 (Change of Variables). If $\mathbf{X}$ is a random vector with the joint PDF of $f_{\mathbf{X}}(x)$ and G is an invertible and continuously differentiable, then the joint PDF of $\mathbf{Y} = G(\mathbf{X})$ is written as follows where $\mathbf{X} = H(\mathbf{Y})$.

$$f_{\mathbf{Y}}(y) = f_{\mathbf{X}}\left(H(y)\right) \cdot \left|\det J_H(y)\right|$$

Aside: In optimal transport, it is often easier to directly utilize H (the inverse) than to find the map G and to then invert to recover the desired density.

To get some practice with this formula, we will work through two examples which show up quite a lot in practice.

Example 8.29 In data assimilation and signal processing, we often have our state vector that we are interested in estimating via time-series data. This forecasting map is often assumed to be linear for not only theoretical benefits of being able to write the density down explicitly, but for its physical meaning in the context of the passage of the signal.

If we assume $\mathbf{Y} = A\mathbf{X} + b$, where $A \in \mathbb{M}_{n\times n}$ and $b \in \mathbb{R}^n$, what is the density of $\mathbf{Y}$? ▪

Proof. We substitute into the change-of-variables formula as follows.

$$\begin{aligned} f_{\mathbf{Y}}(y) &= f_{\mathbf{X}}\left(H(y)\right) \cdot \left|\det J_H(y)\right| && && (8.40)\\ &= f_{\mathbf{X}}\left(A^{-1}\left(y-b\right)\right) \cdot \left|\det J_H(y)\right| && \left(H(\mathbf{Y}) = A^{-1}\left(\mathbf{Y}-b\right)\right) && (8.41)\\ &= \frac{f_{\mathbf{X}}(A^{-1}\left(y-b\right))}{\left|\det A\right|} && \left(\left|\det J_H(y)\right| = \frac{1}{\left|\det A\right|}\right) && (8.42) \end{aligned}$$

■

Example 8.30 If $X, Y \sim \text{NORMAL}(0,1)$ independently, and (R,Θ) are the polar coordinates of (X,Y), find the joint density of (R,Θ). ▪

Proof. We can write down the inverse mappings component-wise as $x = r\cos(\theta)$, $y = r\sin(\theta)$. We first write down the joint density of (X,Y) by recalling the standard normal density.

$$\begin{aligned} f_{X,Y}(x,y) = f_X(x)\cdot f_Y(y) &= \left(\frac{1}{\sqrt{2\pi}}\cdot\exp\left(\frac{-x^2}{2}\right)\right)\cdot\left(\frac{1}{\sqrt{2\pi}}\cdot\exp\left(\frac{-y^2}{2}\right)\right) && (8.43)\\ &= \frac{1}{2\pi}\cdot\exp\left(\frac{-(x^2+y^2)}{2}\right) && (8.44)\\ &= \frac{1}{2\pi}\cdot\exp\left(\frac{-r^2}{2}\right) && (8.45) \end{aligned}$$

We now need to compute the Jacobian of this mapping via the following.

$$\left|\det J\right| = \left|\det\begin{bmatrix}\frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta}\\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta}\end{bmatrix}\right| = \left|\det\begin{bmatrix}\cos(\theta) & -r\sin(\theta)\\ \sin(\theta) & r\cos(\theta)\end{bmatrix}\right| = |r| \tag{8.46}$$

Thus, the joint distribution of (R,Θ) is as follows for $r \geq 0$.

$$f_{R,\Theta}(r,\theta) = \frac{1}{2\pi}\cdot r\exp\left(\frac{-r^2}{2}\right) \equiv f_\Theta(\theta)\cdot f_R(r) \tag{8.47}$$

It turns out that $\Theta \sim \text{UNIFORM}[-\pi,\pi]$ and R is distributed according to the Rayleigh distribution, which yields that (R,Θ) are independent as expected due to being able to factor the distributions. ■

9. Applications of Random Vectors

9.1 Wiener Filters

Agenda:
1. Wiener Filters
2. Gaussian Random Vectors

We begin the lecture by considering the problem of prediction of one random vector from another, in which there are two complementary scenarios but can be mathematically formulated the same.

1. **Prediction of Future**: In the scientific world, we are interested in a certain phenomenon that associates response $\mathbf{Y}$ to input $\mathbf{X}$, and how this can be used for future input $\mathbf{X}$ to obtain predicted values of $\mathbf{Y}$.



2. **Estimation of Past**: In the engineering world, we are given an observed output $\mathbf{Y}$ from a noisy channel and an un-observed signal $\mathbf{X}$ that it came from, and we are interested in estimating the original input $\mathbf{X}$ from the noisy output $\mathbf{Y}$.



We will focus on the second problem for this lecture, but both problems are intimately related in the fact that we are attempting to estimate one vector from another. Thus, we are interested in determining the best estimate $\hat{\mathbf{X}} = g(\mathbf{Y}) \approx \mathbf{X}$.

Aside: This is the framework that can be used to set up inverse problems and data assimilation, the core idea being assimilating data into a probabilistic system.

Before we write down the mathematical statement, we first need to **a**) determine what best means in this context, and **b**) determine what class of functions g can be.

For **a**), the most popular measure of goodness/accuracy of $\hat{\mathbf{X}}$ is the mean squared error (MSE), whose formula is written below. Recall that $\|\cdot\|_2$ refers to the $\mathscr{L}^2$-norm.

$$\text{MSE}(\hat{\mathbf{X}}) = \mathbb{E}[\|\mathbf{X} - \hat{\mathbf{X}}\|_2^2] \tag{9.1}$$

For **b**), we will limit the search space by restricting the class of g to affine functions for simplicity, since we can write down closed-form formulas for how random vectors interact with matrices. Note that if $b = 0$, then this would be the class of linear functions.

$$g(y) = Ay + b \tag{9.2}$$

Aside: This is known as an **admissible** class of functions, and is common terminology in optimization.

With these two clarifications, we are ready to define the problem. Note that we have said *nothing* about the distribution of $\mathbf{X}$ or $\mathbf{Y}$. For those unfamiliar with the notation, the argmin finds the argument x^* that minimized the objective function, equivalently that $f(x^*) \leq f(x)$ for all $x \in \mathscr{D}$ (the search space).

Definition 9.1 The solution to the following problem is known as the **Wiener filter**, or the linear minimum-MSE estimator.

$$g^* = \arg\min_{g \in \mathcal{G}} \mathbb{E}[\|\mathbf{X} - g(\mathbf{Y})\|_2^2] \qquad \mathcal{G} = \{g(y) = Ay + b \mid A \in \mathbb{M}_{n\times n}, b \in \mathbb{R}^n\}$$

Alternatively, one could solve the "unconstrained" optimization problem by not restricting the class of g yet, and solve the minimization problem more generally. This yields a function known as the **regression function** as follows.

Aside: This is the foundational equation to all of statistical learning, which is taught in ACM 158 by Professor Zuev.

$$g^*(y) = \mathbb{E}[X \mid Y = y] \tag{9.3}$$

In parametric regression, we will estimate the parameters of an assumed functional through the data. In non-parametric regression, we will directly attempt to learn $g^*(y)$ through the data. To solve for the Wiener filter, let us directly plug into the MSE formula with the generic affine functional.

$$\begin{aligned}
\mathbb{E}[\|\mathbf{X} - g(\mathbf{Y})\|_2^2] &= \mathbb{E}[\|\mathbf{X} - (A\mathbf{Y} + b)\|_2^2] && (9.4)\\
&= \mathbb{E}[\|(\mathbf{X} - \mu_\mathbf{X}) - (A\mathbf{Y} - A\mu_\mathbf{Y} + b) + \mu_\mathbf{X} - A\mu_\mathbf{Y}\|_2^2] && (9.5)\\
&= \mathbb{E}[\|(\mathbf{X} - \mu_\mathbf{X}) - A(\mathbf{Y} - \mu_\mathbf{Y}) + (\mu_\mathbf{X} - A\mu_\mathbf{Y} - b)\|_2^2] && (9.6)
\end{aligned}$$

Before we handle this directly, we will prove the following lemma.

Lemma 9.2 If $\mathbf{Z}$ is a random vector such that $\mathbb{E}[\mathbf{Z}] = 0$ and c is non-random, then.

$$\mathbb{E}[\|\mathbf{Z} + c\|_2^2] = \mathbb{E}[\|\mathbf{Z}\|^2] + \|c\|^2$$

Proof. This can be done by noting the equivalence of the inner product to the norm as follows.

Aside: This relation of inner products to norms will be essential when working with expectations of norms.

$$\begin{aligned}
\mathbb{E}[\|\mathbf{Z} + c\|_2^2] &= \mathbb{E}[\langle \mathbf{Z} + c, \mathbf{Z} + c\rangle] && (9.7)\\
&= \mathbb{E}[\mathbf{Z}^\top\mathbf{Z} + \mathbf{Z}^\top c + c^\top\mathbf{Z} + c^\top c] && (9.8)\\
&= \mathbb{E}[\mathbf{Z}^\top\mathbf{Z}] + 0 + 0 + \mathbb{E}[c^\top c] && (9.9)\\
&= \mathbb{E}[\|\mathbf{Z}\|^2] + \|c\|^2 && (9.10)
\end{aligned}$$

■

Using this lemma, we can clearly see a random quantity $\mathbf{Z}$ such that $\mathbb{E}[\mathbf{Z}] = 0$ and the non-random component.

$$\begin{aligned}
&\mathbb{E}[\|\underbrace{(\mathbf{X} - \mu_\mathbf{X}) - A(\mathbf{Y} - \mu_\mathbf{Y})}_{\mathbf{Z}} + \underbrace{(\mu_\mathbf{X} - A\mu_\mathbf{Y} - b)}_{c}\|_2^2] && (9.11)\\
&= \mathbb{E}[\|(\mathbf{X} - \mu_\mathbf{X}) - A(\mathbf{Y} - \mu_\mathbf{Y})\|^2] + \|\mu_\mathbf{X} - A\mu_\mathbf{Y} - b\|^2 && (9.12)
\end{aligned}$$

Our goal is to find A and b such that Equation 9.12 is as small as possible. Both terms depend on A, but b appears only in the second term. Therefore, regardless of the value of A, the optimal value of b is such that the following is true.

Aside: Another useful trick in optimization is recognizing when your objective has components that are not a function of what you are optimizing over. Think of this as differentiating a constant.

$$\mu_\mathbf{X} - A\mu_\mathbf{Y} - b = 0 \implies b = \mu_\mathbf{X} - A\mu_\mathbf{Y} \tag{9.13}$$

This now reduces the optimization problem to minimizing the first expectation over A.

$$g^* = \arg\min_{A \in \mathbb{M}_{n\times n}} \mathbb{E}[\|(\mathbf{X} - \mu_\mathbf{X}) - A(\mathbf{Y} - \mu_\mathbf{Y})\|^2] \tag{9.14}$$

We will now prove another lemma to now justify the choice of A.

Lemma 9.3 Let $\mathbf{U}$ and $\mathbf{V}$ be n-dimensional zero-mean random vectors. Then A^* solves the following optimization problem

$$A^* = \arg\min_{A \in \mathbb{M}_{n\times n}} \mathbb{E}[\|\mathbf{U} - A\mathbf{V}\|_2^2] \tag{9.15}$$

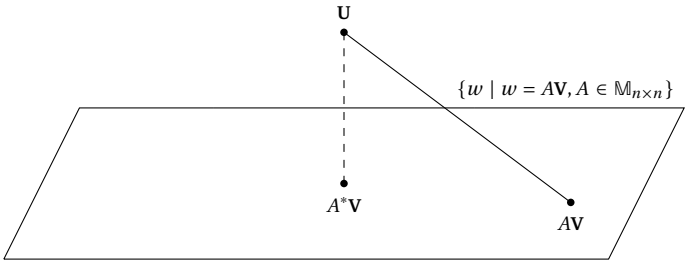
if and only if the following is true for all $B \in \mathbb{M}_{n\times n}$.

$$\mathbb{E}[(B\mathbf{V})^\top \cdot (\mathbf{U} - A^*\mathbf{V})] = 0 \tag{9.16}$$

Proof. We will only consider the backward direction, as this is the easier of the two and the most relevant. To prove the backward direction, we need to show that the following holds true for all $B \in \mathbb{M}_{n\times n}$, as this shows that A^* is the true minimizer .

$$\begin{aligned}
\mathbb{E}[\|\mathbf{U} - B\mathbf{V}\|^2] &\geq \mathbb{E}[\|\mathbf{U} - A^*\mathbf{V}\|^2] && (9.17)\\
&= \mathbb{E}[\|\mathbf{U} - A^*\mathbf{V} + (A^* - B)\mathbf{V}\|^2] && (9.18)\\
&= \mathbb{E}[\|\mathbf{U} - A^*\mathbf{V}\|^2] + 2 \cdot \mathbb{E}[((A^* - B)\mathbf{V})^\top (\mathbf{U} - A^*\mathbf{V})] && (9.19)\\
&\quad + \mathbb{E}[\|(A^* - B)\mathbf{V}\|^2]\\
&= \mathbb{E}[\|\mathbf{U} - A^*\mathbf{V}\|^2] + \mathbb{E}[\|(A^* - B)\mathbf{V}\|^2] && (9.20)\\
&\geq \mathbb{E}[\|\mathbf{U} - A^*\mathbf{V}\|^2] && (9.21)
\end{aligned}$$

In Equation 9.20, we use the assumption of the proof to show that the second component vanishes, which yields desired bound. Thus, the proof of the backwards direction is complete. For a more geometric intuition, consider the following visual.



For a fixed, non-random $\mathbf{U}, \mathbf{V}$, we are looking for the optimal choice of A. From linear algebra, we know that A^* can only be a solution if $\mathbf{U} - A^*\mathbf{V}$ is perpendicular to the plane, which then entails the following condition for all $B \in \mathbb{M}_{n\times n}$.

$$(\mathbf{U} - A^*\mathbf{V}) \perp \text{plane} \Leftrightarrow (\mathbf{U} - A^*\mathbf{V}) \perp B\mathbf{V} \Leftrightarrow (B\mathbf{V})^\top(\mathbf{U} - A^*\mathbf{V}) = 0 \tag{9.22}$$

Thus, the optimality of the matrix A^* is shown. ■

Recall the optimization problem we seek to solve through Equation 9.14. Applying this lemma with $U = \mathbf{X} - \mu_\mathbf{X}$ and $V = \mathbf{Y} - \mu_\mathbf{Y}$, we have that A^* is the minimizer if and only if the following holds for all $B \in \mathbb{M}_{n\times n}$.

$$\begin{aligned}
0 &= \mathbb{E}[(B(\mathbf{Y} - \mu_\mathbf{Y}))^\top \cdot ((\mathbf{X} - \mu_\mathbf{X}) - A^*(\mathbf{Y} - \mu_\mathbf{Y}))] && (9.23)\\
&= \mathbb{E}[\text{tr}((B(\mathbf{Y} - \mu_\mathbf{Y}))^\top \cdot ((\mathbf{X} - \mu_\mathbf{X}) - A^*(\mathbf{Y} - \mu_\mathbf{Y})))] && (9.24)\\
&= \mathbb{E}[\text{tr}(((\mathbf{X} - \mu_\mathbf{X}) - A^*(\mathbf{Y} - \mu_\mathbf{Y})) \cdot (B(\mathbf{Y} - \mu_\mathbf{Y}))^\top)] \quad (\text{tr}(AB) = \text{tr}(BA)) && (9.25)
\end{aligned}$$

Aside: This trick ($\mathbb{E}[\text{tr}(C)] = \text{tr}(\mathbb{E}[C])$) is very common in statistical learning, as we can often have a lot of enjoyed benefits from the trace operator.

Then using the commutativity of the operators, we can split it to exactly recover the covariance matrix definition.

$$\begin{aligned} 0 &= \text{tr}(\mathbb{E}[((\mathbf{X}-\mu_{\mathbf{X}}) - A^*(\mathbf{Y}-\mu_{\mathbf{Y}})) \cdot (B(\mathbf{Y}-\mu_{\mathbf{Y}}))^\top]) && (9.26)\\ &= \text{tr}(\mathbb{E}[((\mathbf{X}-\mu_{\mathbf{X}}) \cdot (\mathbf{Y}-\mu_{\mathbf{Y}})^\top - A^* \cdot (\mathbf{Y}-\mu_{\mathbf{Y}})) \cdot (\mathbf{Y}-\mu_{\mathbf{Y}})^\top \cdot B^\top]) && (9.27)\\ &= \text{tr}((\Sigma_{\mathbf{X},\mathbf{Y}} - A^*\Sigma_{\mathbf{Y}}) \cdot B^\top) && (9.28) \end{aligned}$$

Since this equality must hold for all matrices $B \in \mathbb{M}_{n\times n}$, then this implies the following.

$$A^*\Sigma_{\mathbf{Y}} = \Sigma_{\mathbf{X},\mathbf{Y}} \tag{9.29}$$

From this and Equation 9.13, we can invert to obtain the Wiener filter.

Theorem 9.4 (Wiener Filter). If $\Sigma_{\mathbf{Y}}$ is non-singular, then the Wiener filter is as follows.

$$g(\mathbf{Y}) = \Sigma_{\mathbf{X},\mathbf{Y}}\Sigma_{\mathbf{Y}}^{-1} \cdot (\mathbf{Y}-\mu_{\mathbf{Y}}) + \mu_{\mathbf{X}}$$

Aside: The mean and covariance matrices are typically empirically estimated from the data.

If $\Sigma_{\mathbf{Y}}$ is singular, one can use the pseudo-inverse to still determine the Wiener filter. We will now consider some edge cases depending on the strength of the noise.

1. **Small Noise Limit**: If there is no noise, then $\mathbf{X} = \mathbf{Y}$. Thus, the Wiener filter reduces to the following.

$$g(\mathbf{Y}) = \Sigma_{\mathbf{Y},\mathbf{Y}}\Sigma_{\mathbf{Y}}^{-1} \cdot (\mathbf{Y}-\mu_{\mathbf{Y}}) + \mu_{\mathbf{Y}} = I \cdot (\mathbf{Y}-\mu_{\mathbf{Y}}) + \mu_{\mathbf{Y}} = \mathbf{Y} \tag{9.30}$$

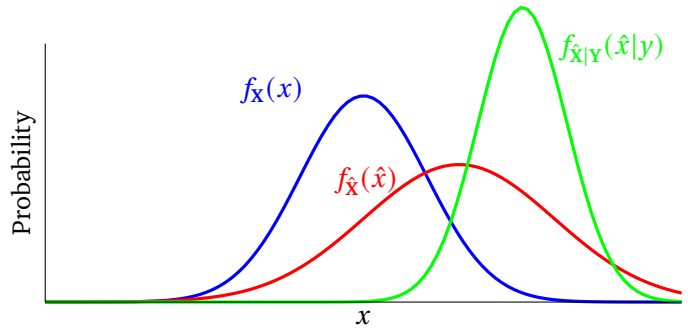
2. **Large Noise Limit**: If the noise is so large that $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated, then $\Sigma_{\mathbf{X},\mathbf{Y}} = 0$, which means that the filter becomes the following.

$$g(\mathbf{Y}) = 0 \cdot \Sigma_{\mathbf{Y}}^{-1} \cdot (\mathbf{Y}-\mu_{\mathbf{Y}}) + \mu_{\mathbf{X}} = \mu_{\mathbf{X}} \tag{9.31}$$

   From this, since the output does not change with respect to input, then knowing $\mathbf{Y}$ does not inform your guess about $\mathbf{X}$. Thus, our best guess can only be the expected value of $\mathbf{X}$.

The Wiener filter can be generalized to the **Kalman** filter in which we assimilate multiple observations, and we attempt to determine what the final probability distribution looks like. Consider the snapshot below, where the probability distribution gets "updated" as the data is assimilated.

Aside: ACM 154 a class on inverse problems and data assimilation that studies these filters extensively.



Of course, there is a tradeoff in the complexity if we complicate the initial distribution or the structure of our data, which is why this is an active field of research.

9.2 Gaussian Random Vectors

A particularly useful subset of random vectors are those whose joint distribution is *Gaussian*.

Definition 9.5 A random vector $X = (X_1, \cdots, X_n)^\top$ is said to be **Gaussian** or normal random vector if its components are jointly normally distributed.

Aside: This can be extended to the infinite-dimensional setting of Gaussian processes, which are prominent in the fields of machine learning and kernel methods.

We recall that for $\mathbf{X}$ to be normally distributed means that there exists matrix A and μ such that the following is true.

$$\mathbf{X} = A\mathbf{Z} + \mu \tag{9.32}$$

Here, $A \in \mathbb{M}_{n\times m}$, $\mathbf{Z} = (Z_1, \cdots, Z_m)^\top$ with $Z_i \sim \text{NORMAL}(0,1)$ i.i.d and $\mu \in \mathbb{R}^n$. In LECTURE 4, we derived the MGF for $\mathbf{X}$ as follows for $t = (t_1, \cdots, t_n)^\top \in \mathbb{R}^n$.

$$M_{\mathbf{X}}(t) = \exp\left(\sum_{i=1}^n t_i\mu_i + \frac{1}{2}\sum_{i,j=1}^n t_i t_j \cdot \text{Cov}(X_i, X_j)\right) \tag{9.33}$$

$$= \exp\left(t^\top \mu + \frac{1}{2} t^\top \Sigma_{\mathbf{X}} t\right) \tag{9.34}$$

It is then clear that the MGF of $\mathbf{X}$ is entirely determined by the mean vector μ and covariance matrix $\Sigma_{\mathbf{X}}$. One can then determine the expectation and covariance of $\mathbf{X}$ using the properties of the covariance matrix discussed in LECTURE 9 as follows.

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}[A\mathbf{Z} + \mu] = A \cdot \mathbb{E}[\mathbf{Z}] + \mu = \mu \tag{9.35}$$

$$\text{Cov}(\mathbf{X}) = \text{Cov}[A\mathbf{Z} + \mu] = A \cdot \text{Cov}(\mathbf{Z}) \cdot A^\top = AA^\top \equiv \Sigma \tag{9.36}$$

Therefore if $\mathbf{X}$ is a Gaussian vector with mean vector μ and covariance matrix Σ, we will say that $\mathbf{X} \sim \text{NORMAL}(\mu, \Sigma)$. Similar to how we showed the effect of linear transformations on random vectors, we will explicitly show it for Gaussian random vectors as well.

Example 9.6 If $\mathbf{X} \sim \text{NORMAL}(\mu_{\mathbf{X}}, \Sigma_{\mathbf{X}})$ and $\mathbf{Y} = B\mathbf{X} + b$, find the distribution of $\mathbf{Y}$. ▪

Aside: This preservation is an important property when discussing normalizing flows, where one repeatedly performs mappings to go from a reference to target distribution.

Proof. One can immediately note that linear mapping will **preserve** Gaussianity, meaning that $\mathbf{Y}$ will also be normally distributed. This can be proven by decomposing $\mathbf{X}$ into the standard normal vector.

$$\mathbf{Y} = B\mathbf{X} + b = B(A\mathbf{Z} + \mu_{\mathbf{X}}) + b \tag{9.37}$$

$$= (BA) \cdot \mathbf{Z} + (B\mu_{\mathbf{X}} + b) \tag{9.38}$$

$$= D\mathbf{Z} + B\mu_{\mathbf{X}} + b \tag{9.39}$$

It is then sufficient to determine the mean vector and covariance matrix.

$$\mu_{\mathbf{Y}} = \mathbb{E}[D\mathbf{Z} + B\mu_{\mathbf{X}} + b] = B\mu_{\mathbf{X}} + b \tag{9.40}$$

$$\Sigma_{\mathbf{Y}} = \text{Cov}(\mathbf{Y}) \tag{9.41}$$

$$= DD^\top \tag{9.42}$$

$$= (BA) \cdot (BA)^\top \tag{9.43}$$

$$= B \cdot (AA^\top) \cdot B^\top \tag{9.44}$$

$$= B\Sigma B^\top \tag{9.45}$$

Thus, $\mathbf{Y} \sim \text{NORMAL}(B\mu + b, B\Sigma B^\top)$. ■

The next natural question is if we can say anything about the *joint probability density* of the components $X_1, \ldots, X_n$ of the random vector $\mathbf{X} \sim \text{NORMAL}(\mu, \Sigma)$. Luckily, we can use the transformation of random variables technique we discussed in LECTURE 9 to determine the density explicitly.

We explicitly derived in LECTURE 9 that if $\mathbf{X} = A\mathbf{Z} + \mu$, where $A \in \mathbb{M}_{n\times n}$ and $\mu \in \mathbb{R}^n$, then the density of $\mathbf{X}$ is as follows. Note that we must assume that A is both *square* and *non-singular* for this to make sense.

$$f_{\mathbf{X}}(x) = \frac{f_{\mathbf{Z}}(A^{-1}(\mathbf{X} - \mu))}{|\det A|} \tag{9.46}$$

It now suffices to determine the joint PDF $f_{\mathbf{Z}}(z)$, which can be factored since the components of $\mathbf{Z}$ are $Z_i \sim \text{NORMAL}(0, 1)$ i.i.d.

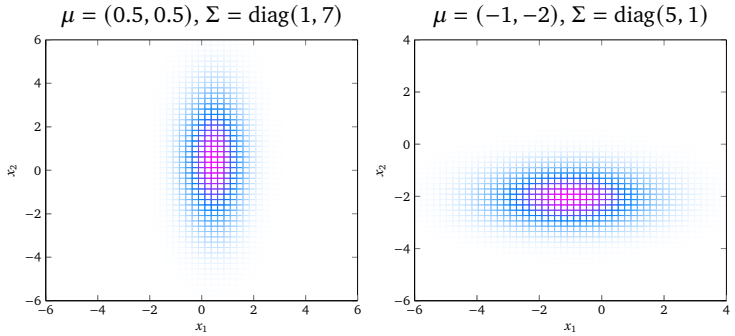
$$\begin{aligned}
f_{\mathbf{Z}}(z) &= \prod_{i=1}^{n} f_{Z_i}(z_i) && (9.47)\\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \cdot \exp\left(\frac{-z_i^2}{2}\right) && (9.48)\\
&= \frac{1}{(2\pi)^{n/2}} \cdot \exp\left(\frac{-1}{2} \cdot \sum_{i=1}^{n} z_i^2\right) && (9.49)\\
&= \frac{1}{(2\pi)^{n/2}} \cdot \exp\left(\frac{-z^\top z}{2}\right) && (9.50)\\
f_{\mathbf{X}}(x) &= \frac{1}{|\det A| \cdot (2\pi)^{n/2}} \cdot \exp\left(\frac{-(A^{-1}(\mathbf{X} - \mu))^\top \cdot (A^{-1}(\mathbf{X} - \mu))}{2}\right) && (9.51)\\
&= \frac{1}{|\det A| \cdot (2\pi)^{n/2}} \cdot \exp\left(\frac{-(\mathbf{X} - \mu)^\top \cdot (AA^\top)^{-1} \cdot (\mathbf{X} - \mu)}{2}\right) && (9.52)
\end{aligned}$$

Aside: You can see the resemblance to the single variable Gaussian density, which should help you remember this formula.

We would now like to express everything in terms of $\Sigma = AA^\top$. Taking the determinant of both sides and using the fact that $\det(A) = \det(A^\top)$, we have that $\det \Sigma = (\det A)^2$. Thus, the final form of the density becomes the following.

$$f_{\mathbf{X}}(x) = \frac{1}{|\det \Sigma|^{1/2} \cdot (2\pi)^{n/2}} \cdot \exp\left(\frac{-(x - \mu)^\top \cdot \Sigma^{-1} \cdot (x - \mu)}{2}\right) \tag{9.53}$$

To paint a more intuitive picture of what this density looks like, consider the following contour plots of a 2-dimensional multivariate Gaussian density and the effects of the parameters on their shape. For those unfamiliar with contour plots, the pink region corresponds to that with the most concentrated density, and the lighter colors indicate less density.



Aside: If we did not have a diagonal covariance, you can imagine the ellipse as being rotated based on the columns of the covariance matrix.

One can see the effects fairly clearly of changing the mean, as this is where the density concentrates the most. The covariance indicates the direction of the spread of the Gaussian, as seen in the two plots above where one of the plots has a much larger variance in the x_2 direction while the other in the x_1 direction. Understanding these visualizations intuitively will help us understand the effects of changing parameters without having to write out the density.

There are a lot of moving parts to this density formula, but we made two critical assumptions about the matrix A.

1. A was non-singular.
2. A was square.

In mathematics, it is of interest to take formulae and test these edge cases, as these often reveal insight into their applications. We will see if we can weaken these conditions, noting that if A is singular, then Σ is singular, and then Equation 9.53 does not makes sense, since the denominator becomes zero due to $\det \Sigma = 0$. If fact, it can be shown that in this case, the joint PDF of $X_1, \ldots, X_n$ does not exist and so $X_1, \ldots, X_n$ are not jointly continuous (although they are jointly normal).

If A is a non-square $n \times m$ matrix ($n \neq m$) and $\Sigma = AA^T$ is $n \times n$ non-singular, then it can be shown that $X_1, \ldots, X_n$ are jointly continuous and Equation 9.53 is their joint PDF. If, however, $\Sigma = AA^T$ is singular, then Equation 9.53 does not make sense and $X_1, \ldots, X_n$ are not jointly continuous.

As a final note, we discussed that if we wanted to recover the input $\mathbf{X}$ from the noisy data $\mathbf{Y}$, then this is done through the following.

$$g^* = \arg\min_g \mathbb{E}[\|\mathbf{X} - g(\mathbf{Y})\|^2] \tag{9.54}$$

If we restrict the admissible set of mappings to affine (of the form $A\mathbf{Y} + b$), we then recover the Wiener filter given by Equation **??**. In general, the solution of Equation **??** is given by the regression function:

$$g(y) = \mathbb{E}[X \mid Y = y] \tag{9.55}$$

To tie what we have discussed this lecture all together, if $\mathbf{X}$ and $\mathbf{Y}$ are both Gaussian random vectors, it turns out that the regression function is indeed the Wiener filter! Gaussianity is a cornerstone in applied mathematics, and should always be looked for in a problem.

10. Introduction to Stochastic Processes

So far, we have considered individual random variables, random vectors, and sequences of i.i.d random variables. The rest of the course will be dedicated to the most interesting probability models: sequences of dependent random variables, or **stochastic processes**.

Agenda:
1. Stochastic Processes
2. Markov Chains
3. Counting Processes
4. Poisson Processes

10.1 Stochastic Processes

We first formally define a stochastic process as follows.

Definition 10.1 A stochastic process $\{X_t, t \in \mathsf{T}\}$ is a collection of random variables such that for every $t \in \mathsf{T}$, X_t is a random variable whose state in state space $\mathfrak{X}$.

Example 10.2 Consider the various examples of stochastic processes below.

1. **Epidemiology**: One can model how many people have a certain virus per day over a month.
2. **Hydrology**: One can model time-series data of precipitation in a certain region.
3. **Physics**: One can model the movement of a particle on a lattice structure.
4. **Mathematical Finance**: One can model the price of a continuously monitored stock.

▪

We are familiar with discrete and continuous random variables, but an additional degree of freedom is incorporated based on the **index set** T. We classify the previous examples based on the structure of their random variables X_t and the index set T.

T \ $\mathfrak{X}$	Discrete	Continuous
Discrete	Ex. 11.2.1	Ex. 11.2.2
Continuous	Ex. 11.2.3	Ex. 11.2.4

Example 10.3 To formalize some more concrete examples of stochastic processes, consider the following.

1. **I.I.D Sequence**: A sequence of i.i.d random variables $\{X_t, t \in \mathbb{N}\}$ is a simple (and very limited from the modeling point of view) example of a discrete-time stochastic process.
2. **Temperature**: Let X_n denote the average temperature on day n. Then $\{X_t, t \in \mathbb{N}\}$ is a discrete-time stochastic process. However, it is clear that X_n and X_m are not independent (the temperature on one day clearly affects the immediate forecast) and the smaller $|n - m|$ is, the stronger the dependence will be.

▪

10.2 Markov Chains

For a general stochastic process $\{X_t, t \in \mathsf{T}\}$, the state X_t at time t may depend on all previous states $\{X_s, s < t\}$. On the other hand it is clear that for many physical phenomena, the dependence between X_t and X_s is small when $|t - s|$ is large.

A Markov chain is a discrete-time stochastic process that models phenomena where X_t depends only on X_{t-1}: given X_{t-1}, X_t is independent from $X_{t-2}, X_{t-3}, \cdots$. We formalize this below.

Definition 10.4 (Markov Process). A stochastic process $\{X_t, t \in \mathsf{T}\}$ where $\mathsf{T} = \mathbb{N} \cup \{0\}$ and $\mathcal{X} = \mathbb{N} \cup \{0\}$ is called a **Markov** chain if

$$\mathbb{P}(X_{n+1} = x \mid X_0 = x_0, \cdots, X_n = x_n) = \mathbb{P}(X_{n+1} = x \mid X_n = x_n)$$

Aside: Intuitively, the Markov property states that the future depends on the past only through the present.

For a general discrete-time state process $\{X_n\}$, the joint PMF can be written as

$$p(x_1, \cdots, x_n) = p_{X_1}(x_1) \cdot p_{X_2 \mid X_1}(x_2 \mid x_1) \cdots p_{X_n \mid X_1, X_2, \cdots, X_{n-1}}(x_n \mid x_1, x_2, \cdots, x_{n-1}) \tag{10.1}$$

Due to the Markov property, we can simplify the conditional densities as follows.

$$p(x_1, \cdots, x_n) = p_{X_1}(x_1) \cdot p_{X_2 \mid X_1}(x_2 \mid x_1) \cdots p_{X_n \mid X_{n-1}}(x_n \mid x_{n-1}) \tag{10.2}$$

One can also generalize a Markovian process to the k-th order by forcing X_{n+1} to depend on the previous k states as opposed to traditional process where $k = 1$.

If one cannot pin down a probabilistic equivalence, but only the expected value, we are then able to define a process to encode this exact property.

Definition 10.5 (Martingale). A stochastic process $\{X_t, t \in \mathsf{T}\}$ where $\mathsf{T} = \mathbb{N} \cup \{0\}$ and $\mathcal{X} = \mathbb{N} \cup \{0\}$ is called a **martingale** if

$$\mathbb{E}(X_{n+1} \mid X_0, \cdots, X_n) = X_n$$

Martingales have an extensive list of applications, namely in mathematical finance. It should be noted that there are martingales that are not Markov chains, and vice versa.

In some sense, a Markov chain can be viewed as the simplest way to induce dependence into a sequence of i.i.d. random variables. But even this weak Markovian dependence makes the theory of Markov chains rich and complex. Some of the many applications are

- modeling asset prices and market crashes
- speech recognition (hidden Markov models)
- biological modeling
- PageRank Algorithm
- Bayesian Statistics & MCMC methods

Aside: For more details, Professor Owhadi teaches a wonderful class (ACM 216) on the theory and applications of Markov chains.

For the rest of the course, we will study

1. Poisson processes, a specific type of counting process (can be interpreted as a continuous-time Markov chain)
2. Brownian motion (aka Weiner process)
3. Characterizing general stochastic processes

10.3 Counting Processes

The most natural process one can think of is the number of events that have happened by some time t. We can formally define such a process as follows.

Definition 10.6 (Counting Process). A stochastic process $\{N_t, t \geq 0\}$ is said to be a **counting process** if

1. $N_t \geq 0$
2. N_t is an integer
3. If $s \leq t$, then $N_s \leq N_t$

Intuitively, N_t represents the number of events that happen before by time t and $N_t - N_s$ represents the number of events that happen in the interval $(s, t]$.

Example 10.7 Some examples of such processes are

a) $N_t = \#$ persons who enter a store **by** time t

b) $N_t = \#$ emails you received **by** time t

c) $N_t = \#$ people born **by** time t ▪

Aside: Note that the definitions replacing **by** with **at** are not counting processes.

As you can imagine, general counting processes are far too general to be amenable for an analytical treatment. To say something meaningful, we need to make some assumptions.

Definition 10.8 A counting process $\{N_t, t \geq 0\}$ is said to have **independent** increments if the number of events that occur in disjoint intervals are independent. Equivalently, for any set of times $t_0 < t_1 < t_2 < \cdots < t_n$,

$$\{N_{t_1} - N_{t_0}, N_{t_2} - N_{t_1}, \cdots, N_{t_n} - N_{t_{n-1}}\} \text{ are mutually independent}$$

We check if such an assumption is reasonable on the example processes proposed in Example 11.3.

(a) **Yes**, since the current amount of customers has no effect on the future amount.
(b) **Yes**, since the current amount of emails has no effect on the future amount.
(c) **No**, since many people born initially will likely imply many people born later.

Definition 10.9 A counting process $\{N_t, t \geq 0\}$ is said to have **stationary** increments if the number of events that occur in any interval of time $(s, s + t]$ depends only on the length of the interval t. Equivalently, for all t_1, t_2, s,

$$N_{t_2+s} - N_{t_1+s} \sim N_{t_2} - N_{t_1} \text{ (in distribution)}$$

We once again check if such an assumption is reasonable on the example processes proposed in Example 11.3.

(a) **No** if rush hour; **Yes** if not rush hour
(b) **No** if you communicate with colleagues; **Yes** if you do not communicate with colleagues
(c) **No** large time horizon; **Yes** small time horizon

10.4 Poisson Processes

While counting processes are a very natural way of describing a process, we have not said anything about how **frequent** such events happen. Thus, we can impose a distribution on the number of events that occur in a particular interval, which leads to the definition of a Poisson process.

Aside: Renewal processes are Poisson processes with a general waiting time distribution.

Definition 10.10 (Poisson Process). The counting process $\{N_t, t \geq 0\}$ is called a **Poisson process** with rate $\lambda > 0$ if,

(1) $N_0 = 0$
(2) $\{N_t\}$ has stationary and independent increments
(3) For all $s, t \geq 0$, $N_{t+s} - N_s \sim \text{POISSON}(\lambda t)$. Equivalently, for all $n \in \mathbb{N} \cup \{0\}$,

$$\mathbb{P}(N_{t+s} - N_s = n) = \exp(-\lambda t) \cdot \frac{(\lambda t)^n}{n!}$$

In particular, letting $s = 0$ for $(\mathbf{3})$ yields the following. Thus, λ is often referred to as the **rate** of the process.

$$N_t \sim \text{POISSON}(\lambda t) \implies \mathbb{E}[N_t] = \lambda t \implies \lambda = \frac{\mathbb{E}[N_t]}{t} \tag{10.3}$$

Suppose we want to model the phenomena in Example 11.3. To decide whether the Poisson process is a reasonable model, we need to check whether the assumptions $(\mathbf{1}), (\mathbf{2}), (\mathbf{3})$ are reasonable. The first two are fairly straightforward as

1. $(\mathbf{1})$ we can simply assume we start counting at $t = 0$
2. $(\mathbf{2})$ we can verify through the physics of the phenomena

The assumption that is not immediately clear to verify as reasonable is $(\mathbf{3})$. It turns out that if $(\mathbf{1})$ and $(\mathbf{2})$ are true, then $(\mathbf{3})$ is equivalent to $(\mathbf{3})^*$.

Aside: Recall that $\mathfrak{o}(h)$ is the following class of functions.

$$\mathfrak{o}(h) = \left\{ f \;\middle|\; \lim_{h \to 0} \frac{f(h)}{h} = 0 \right\}$$

As an example, $x^2 = \mathfrak{o}(x)$.

$$(\mathbf{3})^* = \begin{cases} \mathbb{P}(N_h = 1) = \lambda h + \mathfrak{o}(h), & \text{as } h \to 0 \\ \mathbb{P}(N_h \geq 2) = \mathfrak{o}(h), & \text{as } h \to 0 \end{cases} \tag{10.4}$$

We remark what these probabilities intuitively mean as follows.

- For $h \ll 1$, the probability of an event happening in the interval $(0, h]$ (can be any interval of length h due to stationarity) is proportional to h, the length of the interval.
- For $h \ll 1$, the probability of at least two events happening in the interval $(0, h]$ tends to 0 faster than h itself.

To show the equivalence between $(\mathbf{3})$ and $(\mathbf{3})^*$, we need to show that $(\mathbf{3}) \implies (\mathbf{3})^*$ **and** $(\mathbf{3})^* \implies (\mathbf{3})$. The former is trivial through direct substitution.

$$\begin{aligned} \mathbb{P}(N_h = 1) &= (\lambda h) \cdot \exp(-\lambda h) && (10.5) \\ &= \lambda h \cdot (1 - \lambda h + \mathfrak{o}(h)) && (10.6) \\ &= \lambda h + (-\lambda^2 h^2 + \lambda h \cdot \mathfrak{o}(h)) && (10.7) \\ &= \lambda h + \mathfrak{o}(h) && (10.8) \end{aligned}$$

Aside: We use the Taylor expansion of $\exp(x)$ for $x \ll 1$ as

$$\exp(x) = 1 + x + \mathfrak{o}(x)$$

$$\begin{aligned} \mathbb{P}(N_h \geq 2) &= 1 - \mathbb{P}(N_h = 0) - \mathbb{P}(N_h = 1) && (10.9) \\ &= 1 - \exp(-\lambda h) - (\lambda h + \mathfrak{o}(h)) && (10.10) \\ &= 1 - (1 - \lambda h + \mathfrak{o}(h)) - \lambda h + \mathfrak{o}(h) && (10.11) \\ &= \mathfrak{o}(h) && (10.12) \end{aligned}$$

To show the converse $(3)^* \implies (3)$ is challenging. We seek to show that $N_t \sim \text{POISSON}(\lambda t)$, which, due to the stationarity of increments (2), will imply (3). To do this, consider the PMF of N_{t+h}.

$$\mathbb{P}(N_{t+h} = n) = \sum_{k=0}^{n} \mathbb{P}(N_t = k, N_{t+h} - N_t = n - k) \tag{10.13}$$

Intuitively, this says that we observe k events in the time interval $(0, t]$ and $n - k$ more events in the time interval $(t, t + h]$. Taking this summation over all possible k must yield the total probability that we see n events by time $t + h$. By independent increments, this probability splits as follows.

$$\mathbb{P}(N_{t+h} = n) = \sum_{k=0}^{n} \mathbb{P}(N_t = k) \cdot \mathbb{P}(N_{t+h} - N_t = n - k) \tag{10.14}$$

$$= \sum_{k=0}^{n} \mathbb{P}(N_t = k) \cdot \mathbb{P}(N_h = n - k) \tag{10.15}$$

Using the statements in $(3)^*$, we can split this summation based on k as follows.

$$\mathbb{P}(N_{t+h} = n) = \mathbb{P}(N_t = n) \cdot \mathbb{P}(N_h = 0) + \mathbb{P}(N_t = n - 1) \cdot \mathbb{P}(N_h = 1) + \sum_{k=0}^{n-2} \mathbb{P}(N_t = k) \cdot \mathbb{P}(N_h = n - k) \tag{10.16}$$

The last term contains $\mathbb{P}(...)$, where $n - k \geq 2$ since $k \in [0, n - 2]$. Therefore, each of these probabilities, and thus the entire term, is $\mathfrak{o}(h)$. This simplifies the probability as follows.

$$\mathbb{P}(N_{t+h} = n) = \mathbb{P}(N_t = n) \cdot \mathbb{P}(N_h = 0) + \mathbb{P}(N_t = n - 1) \cdot \mathbb{P}(N_h = 1) + \mathfrak{o}(h) \tag{10.17}$$

$$= \mathbb{P}(N_t = n) \cdot (1 - \lambda h) + \mathbb{P}(N_t = n - 1) \cdot (\lambda h) + \mathfrak{o}(h) \tag{10.18}$$

Rewriting this as a difference equation, we obtain the following.

$$\frac{\mathbb{P}(N_{t+h} = n) - \mathbb{P}(N_t = n)}{h} = -\lambda \cdot \mathbb{P}(N_t = n) + \lambda \cdot \mathbb{P}(N_t = n - 1) + \frac{\mathfrak{o}(h)}{h} \tag{10.19}$$

Letting $h \to 0$, we obtain the following sequence of differential equations (parameterized by n) on the probability $\mathbb{P}(N_t = n)$.

$$\frac{d\mathbb{P}(N_t = n)}{dt} = -\lambda \cdot \mathbb{P}(N_t = n) + \lambda \cdot \mathbb{P}(N_t = n - 1) \tag{10.20}$$

For the case where $n = 0$, we then have the following differential equation with the initial condition that $\mathbb{P}(N_0 = 0) = 1$.

$$\frac{d\mathbb{P}(N_t = 0)}{dt} = -\lambda \cdot \mathbb{P}(N_t = 0) \implies \mathbb{P}(N_t = 0) = \left. \frac{e^{-\lambda t} \cdot (\lambda t)^n}{n!} \right|_{n=0} \tag{10.21}$$

By induction on n, one can show that not only for $n = 0$, but for any n

$$\mathbb{P}(N_t = n) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}. \tag{10.22}$$

Thus, $N_t \sim \text{POISSON}(\lambda t)$. The Poisson process is one of the most widely used counting processes, used when certain events appear to happen at a certain rate but otherwise random occurences with no structure. The Poisson process and its extensions have been used to model **a**) earthquake occurences, **b**) traffic accidents, **c**) email arrivals, and **d**) webpage requests.

Aside: One can prove this by interpreting the Laplace transform as an expectation.

$$g(t) = \mathbb{E}[\exp(-u \cdot N_t)]$$

11. Poisson Processes

Last time we introduced the Poisson process, which is a counting process $\{N_t, t \geq 0\}$, satisfying the following conditions:

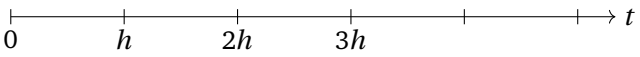
1. $N_0 = 0$
2. $\{N_t\}$ has independent and stationary increments
3. $N_{t+s} - N_s \sim \text{POISSON}(\lambda t)$

In this lecture, we will describe the Poisson process through its limiting process, as well as its associated properties.

Agenda:
1. Limit of Bernoulli Process
2. Interarrival Times
3. Waiting Times

11.1 Limit of Bernoulli Process

Suppose we want to model the occurence of events (N_t) that happen completely at random with rate λ per unit time. If we start counting at time $t = 0$, then we can let $N_0 = 0$. We can then break up the timeline as follows.



Formally, we split the timeline as a **disjoint** union of each interval $\mathcal{I}_k$ of length h.

$$(0, \infty) = \bigsqcup_{k=1}^{\infty} \mathcal{I}_k \qquad \mathcal{I}_k = ((k-1)h, kh) \tag{11.1}$$

Aside: It is often informative to study the limiting behavior of simpler processes to obtain intuition on more complicated ones.

From our intuition of the Poisson process, we can now probabilistically model whether an event occurs in each of the intervals $\mathcal{I}_k$ by an independent Bernoulli random variable $X_k \sim \text{BERNOULLI}(\lambda h)$ defined below.

$$X_k = \begin{cases} 1 & \text{w.p} \quad \lambda h & \text{(event happened in } \mathcal{I}_k\text{)} \\ 0 & \text{w.p} \quad 1 - \lambda h & \text{(event did not happen in } \mathcal{I}_k\text{)} \end{cases} \tag{11.2}$$

Thus, N_t will simply be the sum of these Bernoulli random variables that indicate whether an event happened per unit time interval, which can be directly recognized as a binomial random variable where $n = \lfloor t/h \rfloor$.

$$N_t = \sum_{k=1}^{n} X_k \tag{11.3}$$

If we let the time step $h \to 0$, we directly obtain that $n \to \infty$ and $p \to 0$. Their product, however, yields the desired rate.

$$np \approx \frac{t}{h} \cdot \lambda h = \lambda t \tag{11.4}$$

From LECTURE 2, you will recall these are exactly the limits needed to recover the Poisson distribution. Thus, $N_t \sim \text{POISSON}(\lambda t)$. Furthermore, it is clear that by modeling each event as a Bernoulli random variable on a unit time interval that we immediately obtain independent and stationary increments. In total, $\{N_t\}$ is indeed a Poisson process.

11.2 Interarrival Times

Let $\{N_t, t \geq 0\}$ be a Poisson process with rate λ. Let T_1 denote the time of the first event, and T_n denote the time between the $(n-1)$-th event and n-th event.

Definition 11.1 (Interarrival Times). The sequence $\{T_n, n \in \mathbb{N}\}$ is called the sequence of **interarrival** times.

It is clear that T_n will be a random variable, but it is of interest to characterize its distribution. To do this, we consider the CDF of T_1 as follows. Note that $T_1 > t$ is equivalent to the fact that no events have happened before time t. From this, we have that $T_1 \sim \text{EXPONENTIAL}(\lambda)$.

$$\begin{aligned} \mathbb{P}(T_1 \leq t) &= 1 - \mathbb{P}(T_1 > t) && (11.5)\\ &= 1 - \mathbb{P}(N_t = 0) && (11.6)\\ &= 1 - \exp(-\lambda t) && (11.7) \end{aligned}$$

We can attempt to characterize the CDF of T_2, using conditional expectation to re-express in a more amenable form.

$$\begin{aligned} \mathbb{P}(T_2 \leq t) &= 1 - \mathbb{P}(T_2 > t) && (11.8)\\ &= 1 - \mathbb{E}_{T_1}[\mathbb{P}(T_2 > t \mid T_1)] \qquad \text{(law of total probability)} && (11.9) \end{aligned}$$

We now note that for a fixed event time $T_1 = s$, we can simplify using the properties of the Poisson process. Once again, $T_2 > t$ indicates that the second event has not happened t units after the first event.

$$\begin{aligned} \mathbb{P}(T_2 > t \mid T_1 = s) &= \mathbb{P}(N_{t+s} - N_s = 0 \mid \overbrace{N_s = 1, \{N_r\}_{r\in[0,s)} = 0}^{T_1 = s}) && (11.10)\\ &= \mathbb{P}(N_t = 0) \qquad \text{(independence \& stationarity)} && (11.11)\\ &= \exp(-\lambda t) && (11.12) \end{aligned}$$

Thus, our original probability becomes the following, noting that the expectation over T_1 of a constant simplifies.

$$\mathbb{P}(T_2 \leq t) = 1 - \mathbb{E}_{T_1}[\exp(-\lambda t)] = 1 - \exp(-\lambda t) \tag{11.13}$$

Aside: The i.i.d result on the interarrival times will be useful when simulating a Poisson process.

From this, we have that $T_2 \sim \text{EXPONENTIAL}(\lambda)$ and that $T_1 \perp\!\!\!\perp T_2$ due to independent increments. Thus, we have the following theorem regarding the interarrival times of a Poisson process.

Theorem 11.2 If $\{N_t\}$ is a Poisson process with rate λ, then the interarrival times $T_1, \cdots, T_n \sim \text{EXPONENTIAL}(\lambda)$ i.i.d.

It turns out that the exponential random variable (resp. geometric random variable) is the **only** continuous (resp. discrete) random variable to be **memoryless**.

Definition 11.3 (Memoryless). A non-negative random variable X is said to be **memoryless** if for all $s, t \geq 0$,

$$\mathbb{P}(X > s + t \mid X > t) = \mathbb{P}(X > s)$$

To show that there are no other random variables satisfy it is more complicated and will not be discussed here. The *heuristic* argument is that if one writes down the functional equation in terms of the complementary CDFs, the exponential function is the only choice to satisfy the equation. However, to show that the exponential random variable satisfy the property is trivial, as shown below.

Proof. Using the Bayes' rule, the memoryless property is equivalent to the conditional probability.

$$\mathbb{P}(X > s) = \frac{\mathbb{P}(X > s + t, X > t)}{\mathbb{P}(X > t)} \tag{11.14}$$

$$\mathbb{P}(X > s + t) = \mathbb{P}(X > t) \cdot \mathbb{P}(X > s) \tag{11.15}$$

The exponential random variable does satisfy this equation (therefore, memoryless):

$$\exp(-\lambda \cdot (s + t)) = \exp(-\lambda s) \cdot \exp(-\lambda t) \tag{11.16}$$

■

Intuitively, this can be described by considering an electronic component. The memoryless property indicates that if a component has lived for t hours, the probability it will survive for s **more** hours is the same as the initial probability that it will survive for s hours. Such a property is consistent with the independent and stationary increments assumption made by the Poisson process.

Example 11.4 We will experiment with this memoryless property of a Poisson process N_t with rate λ explicitly.

(a) Consider the following probability that the first event happens later than time t_0.

$$\mathbb{P}(T_1 > t_0) = \exp(-\lambda t_0) \tag{11.17}$$

(b) If we knew that there were no events before t_0, the probability that the event happens later than t_1 is not affected.

$$\mathbb{P}(T_1 > t_1 \mid T_1 > t_0) = \mathbb{P}(T_1 > t_1 - t_0) = \exp(-\lambda(t_1 - t_0)) \tag{11.18}$$

(c) If we now knew that the third event happened at time t_0, the probability that the fourth event happens later than t_1 is not affected since the T_i are independent.

$$\mathbb{P}(T_4 > t_1 - t_0 \mid T_1 + T_2 + T_3 = t_0) = \mathbb{P}(T_4 > t_1 - t_0) = \exp(-\lambda(t_1 - t_0)) \tag{11.19}$$

(d) If we start viewing the process at time t_0, and T is the time of the first event we see, then $T = t_0 + T_1$ due to the probabilistic restart at t_0. Thus, the moments can be computed as follows.

$$\mathbb{E}[T] = t_0 + \frac{1}{\lambda} \qquad \mathbb{V}[T] = \frac{1}{\lambda^2} \tag{11.20}$$

(e) If we start viewing the process at time t_0, and T is the time of the first event we see, then the conditional expectation and variance of T given that the last event occured at time t^* are unchanged from **(d)**.

$$\mathbb{E}[T] = t_0 + \frac{1}{\lambda} \qquad \mathbb{V}[T] = \frac{1}{\lambda^2} \tag{11.21}$$

This is because occurences before and after $t = t_0$ are independent of each other, and information about the last occurence is irrelevant.

▪

11.3 Waiting Times

While the interarrival times are an important statistic about the process, the total time that elapses before an event is of interest as well.

Definition 11.5 The **waiting** time until the n-th event denoted by S_n is the time till the n-th event occurs, where $n \geq 1$ and $S_0 = 0$.

$$S_n = \sum_{i=1}^{n} T_i = S_{n-1} + T_n$$

It is clear that the waiting times are random quantities, but lose the benefit of independence since $S_1 \leq S_2 \leq \cdots S_n$ due to the recursive definition. However, the first two moments are trivial to compute since S_n is an sum of i.i.d random variables.

$$\mathbb{E}[S_n] = \sum_{i=1}^{n} \mathbb{E}[T_i] = \frac{n}{\lambda} \tag{11.22}$$

$$\mathbb{V}[S_n] = \sum_{i=1}^{n} \mathbb{V}[T_i] = \frac{n}{\lambda^2} \tag{11.23}$$

Aside: This supports our intuition with direct proportionality to n and inversely so to λ.

To characterize the distribution of S_n, we attempt to do this directly through the CDF of S_n. Similar to how we related T_i to the counting process N_t, it is clear that $S_n \leq t$ iff $N_t \geq n$.

$$F_{S_n}(t) = \mathbb{P}(S_n \leq t) = \mathbb{P}(N_t \geq n) = \sum_{k=n}^{\infty} \exp(-\lambda t) \cdot \frac{(\lambda t)^k}{k!} \tag{11.24}$$

However, to determine the PDF of S_n, we would need to take the derivative which would be quite messy. Instead, we will directly work with the PDF using the same intuition for a generic PDF.

$$\mathbb{P}\left(t - \frac{h}{2} \leq S_n \leq t + \frac{h}{2}\right) = \mathbb{P}\left(N_{t-\frac{h}{2}} = n-1, N_{t+\frac{h}{2}} - N_{t-\frac{h}{2}} \geq 1\right) \tag{11.25}$$

$$= \mathbb{P}\left(N_{t-\frac{h}{2}} = n-1\right) \cdot \mathbb{P}\left(N_h \geq 1\right) \tag{11.26}$$

$$= \mathbb{P}\left(N_{t-\frac{h}{2}} = n-1\right) \cdot \left(\mathbb{P}\left(N_h = 1\right) + \mathbb{P}\left(N_h \geq 2\right)\right) \tag{11.27}$$

$$= \mathbb{P}\left(N_{t-\frac{h}{2}} = n-1\right) \cdot \left(\lambda h + \mathfrak{o}(h)\right) \tag{11.28}$$

This can now be utilized in the limit definition of the PDF, which yields exactly the Gamma distribution with shape parameter n and scale parameter λ.

$$f_{S_n}(t) = \lim_{h \to 0} \frac{\mathbb{P}\left(t - \frac{h}{2} \leq S_n \leq t + \frac{h}{2}\right)}{h} \tag{11.29}$$

$$= \lim_{h \to 0} \frac{\mathbb{P}\left(N_{t-\frac{h}{2}} = n-1\right) \cdot \left(\lambda h + \mathfrak{o}(h)\right)}{h} \tag{11.30}$$

$$= \lim_{h \to 0} \left(\mathbb{P}\left(N_{t-\frac{h}{2}} = n-1\right) \cdot \lambda + \frac{\mathfrak{o}(h)}{h}\right) \tag{11.31}$$

$$= \mathbb{P}\left(N_t = n-1\right) \cdot \lambda \tag{11.32}$$

$$= \exp(-\lambda t) \cdot \frac{(\lambda t)^{n-1}}{(n-1)!} \cdot \lambda \tag{11.33}$$

Theorem 11.6 If $\{N_t\}$ is a Poisson process with rate λ, then the waiting times $S_1, S_2, \cdots$ have a distribution

$$S_n \sim \text{GAMMA}(n, \lambda)$$

The discussed properties of the interarrival times and waiting times allow us to give yet another definition of the Poisson process that is useful for simulation. As mentioned before, the benefit of the i.i.d property of the interarrival times is that we can easily sample without having to induce any additional dependence between the variables.

Furthermore, we learned how to sample from the exponential distribution using the inverse transform method, although most numerical packages will directly implement this. However, for a more general interarrival distribution, this would be useful.

Algorithm 11.1 Poisson Process Simulation

Input: Rate parameter λ
Initialize: Set $S_0 = 0$
for each event $n = 1, 2, 3, \ldots$ **do**
    Draw $T_n \sim \text{Exp}(\lambda)$ ▷ Exponential random variable with rate λ
    Set $S_n = S_{n-1} + T_n$ ▷ Sum of waiting times
$N_t = \max\{n : S_n \le t\}$ ▷ Events before time t

Aside: The `MATLAB` code for this simulation is on Page 447 of GUBNER.

Warning 11.7 While increments are independent by the Poisson process assumptions,

$$N_{t_2} - N_{t_1} \perp\!\!\!\perp N_{t_1} - N_{t_0} = N_{t_1}$$

the random variables N_{t_1} and N_{t_2} themselves are **not** independent. ▪

The natural question is then **how related** are N_{t_1} and N_{t_2}, which we can measure through the covariance. For this, we assume that $t_1 < t_2$ without loss of generality.

$$\begin{aligned} \text{Cov}(N_{t_1}, N_{t_2}) &= \mathbb{E}[N_{t_1} N_{t_2}] - \mathbb{E}[N_{t_1}] \cdot \mathbb{E}[N_{t_2}] && (11.34) \\ &= \mathbb{E}[N_{t_1} N_{t_2}] - (\lambda t_1) \cdot (\lambda t_2) && (11.35) \end{aligned}$$

To handle the first term, we can re-interpret N_{t_2} in terms of N_{t_1} as follows.

$$\begin{aligned} \mathbb{E}[N_{t_1} N_{t_2}] &= \mathbb{E}[N_{t_1} \cdot (N_{t_2} - N_{t_1} + N_{t_1})] && (11.36) \\ &= \mathbb{E}[N_{t_1} \cdot (N_{t_2} - N_{t_1})] + \mathbb{E}[N_{t_1}^2] && (11.37) \\ &= \mathbb{E}[N_{t_1}] \cdot \mathbb{E}[N_{t_2} - N_{t_1}] + (\mathbb{V}[N_{t_1}] + \mathbb{E}[N_{t_1}]^2) \quad \text{(independence)} && (11.38) \\ &= (\lambda t_1) \cdot \mathbb{E}[N_{t_2 - t_1}] + (\lambda t_1 + (\lambda t_1)^2) \quad \text{(stationarity)} && (11.39) \\ &= (\lambda t_1) \cdot (\lambda \cdot (t_2 - t_1)) + (\lambda t_1 + (\lambda t_1)^2) && (11.40) \\ &= \lambda^2 t_1 t_2 + \lambda t_1 && (11.41) \end{aligned}$$

Aside: This trick of adding 0 is essential in being able to use stationarity and independence.

Substituting into the original equation, we have the following.

$$\text{Cov}(N_{t_1}, N_{t_2}) = \lambda^2 t_1 t_2 + \lambda t_1 - \lambda^2 t_1 t_2 = \lambda t_1 = \mathbb{V}[N_{t_1}] \tag{11.42}$$

More generally, the covariance formula is as follows.

$$\text{Cov}(N_{t_1}, N_{t_2}) = \lambda \cdot \min\{t_1, t_2\} \tag{11.43}$$

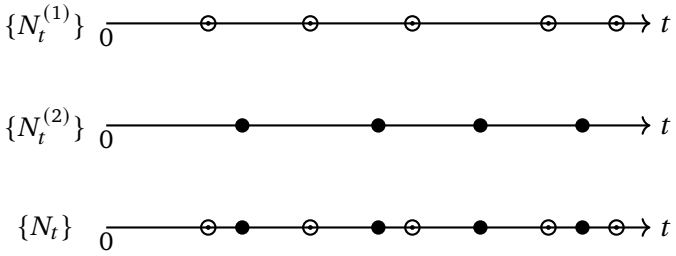
12. More Poisson Processes

Last time we discussed properties of individual Poisson processes. In this lecture, we will learn how to **merge** and **split** them, which are two operations very important in applications.

Agenda:
1. Merging Processes
2. Splitting Process
3. E-Commerce Application
4. Conditional Arrival Times
5. Order Statistics

12.1 Merging Processes

Let $\{N_t^{(1)}\}_{t\geq 0}$ and $\{N_t^{(2)}\}_{t\geq 0}$ be two independent Poisson processes with rates λ_1 and λ_2, respectively. We also enforce **independence** of the processes $\{N_t^{(1)}\} \perp\!\!\!\perp \{N_t^{(2)}\}$, which means that neither of the processes affect each other. It is now of interest to combine these processes with $N_t = N_t^{(1)} + N_t^{(2)}$, as pictured below.



The claim is then that the combined process is also a Poisson process such that $\{N_t\}_{t\geq 0}$ is a **Poisson process** with rate $\lambda_1 + \lambda_2$. We can verify the requisite conditions for a Poisson process on this merged process as follows.

1. $N_0 = N_0^{(1)} + N_0^{(2)} = 0$
2. Independent increments of N_t is an immediate consequence of the independence property of each of the individual processes.
3. Stationarity can be shown through the splitting of the two processes.

$$N_{t+s} - N_s = \left(N_{t+s}^{(1)} - N_s^{(1)}\right) + \left(N_{t+s}^{(2)} - N_s^{(2)}\right) \sim \text{POISSON}((\lambda_1 + \lambda_2)t) \quad (12.1)$$

This construction can be directly generalized to any number of **independent** Poisson processes through the additive property of Poisson random variables.

Theorem 12.1 (Merging Processes). If $\{N_t^{(1)}\}_{t\geq 0}, \{N_t^{(2)}\}_{t\geq 0}, \cdots, \{N_t^{(m)}\}_{t\geq 0}$ are m independent Poisson processes with rates $\lambda_1, \cdots, \lambda_m$. Then, $\{N_t\}_{t\geq 0}$ is a Poisson process with rate λ.

$$N_t = \sum_{i=1}^m N_t^{(i)} \implies N_t \sim \text{POISSON}(\lambda t) \quad \text{where} \quad \lambda = \sum_{i=1}^m \lambda_i$$

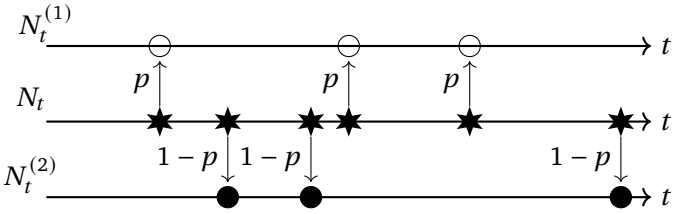
12.2 Splitting Processes

Let $\{N_t\}_{t\geq 0}$ be a Poisson process with rate λ. Suppose that each time an event occurs, it is classified as a type I event with probability p, or a type II event with probability $1-p$, independently of the other events. Formally, we claim that this splitting yields Poisson processes, defined below with $\{N_t^{(1)}\}_{t\geq 0} \perp\!\!\!\perp \{N_t^{(2)}\}_{t\geq 0}$.

$$\{N_t^{(1)}\}_{t\geq 0} \sim \text{POISSON}(\lambda p) \qquad \{N_t^{(2)}\}_{t\geq 0} \sim \text{POISSON}(\lambda(1-p)) \tag{12.2}$$

Aside: It is clear that $N_t = N_t^{(1)} + N_t^{(2)}$.

Once again, consider the following visual of a Poisson process getting split.



We now seek to prove that the resulting stochastic processes are indeed Poisson, which we will verify without loss of generality for $N_t^{(1)}$.

1. $N_0^{(1)} = 0$ since $N_0 = 0$.
2. We can use the independence of increments of $\{N_t\}_{t\geq 0}$ to show the independence of increments of $\{N_t^{(1)}\}_{t\geq 0}$. If one extracts two intervals $\mathcal{I}_1$ and $\mathcal{I}_2$ from the timeline of N_t, and those same intervals $\mathcal{I}_1^{(1)}$ and $\mathcal{I}_2^{(1)}$ from the timeline of $N_t^{(1)}$, it is clear that

   $$\mathcal{I}_1^{(1)} \text{ depends on } \mathcal{I}_1 \qquad \mathcal{I}_2^{(1)} \text{ depends on } \mathcal{I}_2$$

   Thus, if $\mathcal{I}_1^{(1)}$ and $\mathcal{I}_2^{(1)}$ are dependent, so are $\mathcal{I}_1$ and $\mathcal{I}_2$, which is a contradiction. Thus, intervals in our split process will be independent.
3. We can attempt to prove stationarity explicitly, conditioning on how many events occur in our regular process as follows since we know $\{N_t\}_{t\geq 0}$ has stationary increments.

   $$\mathbb{P}(N_{t+s}^{(1)} - N_s^{(1)} = k) = \sum_{n=k}^{\infty} \mathbb{P}(N_{t+s}^{(1)} - N_s^{(1)} = k \mid N_{t+s} - N_s = n) \cdot \mathbb{P}(N_{t+s} - N_s = n) \tag{12.3}$$

   $$= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \cdot \exp(-\lambda t) \cdot \frac{(\lambda t)^n}{n!} \tag{12.4}$$

   Since this probability depends only on the length t of the time interval $(s, s+t]$, the process $\{N_t^{(1)}\}_{t\geq 0}$ has stationary increments.

   While this does depend only on t, the rate of $N_t^{(1)}$ is unclear from the above expression. As such, we can use the alternative definition of the Poisson process by using the total law of probability.

   $$\mathbb{P}(N_h^{(1)} = 1) = \mathbb{P}(N_h^{(1)} = 1 \mid N_h = 1) \cdot \mathbb{P}(N_h = 1) + \mathbb{P}(N_h^{(1)} = 1 \mid N_h \geq 2) \cdot \mathbb{P}(N_h \geq 2) \tag{12.5}$$

   $$= p \cdot (\lambda h + \mathfrak{o}(h)) + \mathbb{P}(N_h^{(1)} = 1 \mid N_h \geq 2) \cdot \mathfrak{o}(h) \tag{12.6}$$

   $$= (\lambda p) \cdot h + \mathfrak{o}(h) \tag{12.7}$$

   $$\mathbb{P}(N_h^{(1)} \geq 2) \leq \mathbb{P}(N_h \geq 2) = \mathfrak{o}(h) \tag{12.8}$$

Thus, $\{N_t^{(1)}\}_{t\geq 0}$ is a Poisson process with rate λp. This can again be generalized for any number of events we wish to split our Poisson process into, but it must be treated with some care to ensure that mutual independence of the class of events holds.

12.3 E-Commerce Application

Suppose we want to sell an item online via eBay or another platform. We make the following assumptions on the process.

1. We receive offers according to a Poisson process with intensity λ.
2. Each offer is modeled as a random variable $X \sim f_X(x)$, $x \geq 0$.
3. Once an offer is received, we can accept it or reject it and wait for the next offer.
4. We incur fees at a rate of c per unit time until the time of the item being sold.

With all this, our goal is to maximize the expected total return R.

$$R = \text{accepted offer} - c \cdot (\text{time to accept}) \tag{12.9}$$

Aside: It can be shown with much effort that the x_0-policy is optimal.

A policy known as the x_0-policy is to accept the first offer that is larger than x_0. We can formalize the equation for R as follows.

$$\mathbb{E}[R(x_0)] = \mathbb{E}[X \mid X > x_0] - c \cdot \mathbb{E}[T]$$

We can formulate the conditional density through Bayes' rule as follows. Here, $\mathcal{I}_{x_0,\infty}(x)$ represents the indicator function on the interval (x_0, ∞).

$$f_{X \mid X>x_0} = \frac{\mathbb{P}(X > x_0 \mid X = x) \cdot f_X(x)}{\mathbb{P}(X > x_0)} = \frac{\mathcal{I}_{x_0,\infty}(x) \cdot f_X(x)}{\mathbb{P}(X > x_0)} \tag{12.10}$$

Thus, the conditional expectation simplifies, where $\overline{F}(x_0) = 1 - \mathbb{P}(X \leq x_0) = 1 - F(x_0)$ is known as the complementary CDF.

$$\mathbb{E}[X \mid X > x_0] = \int_0^\infty x \cdot \frac{\mathcal{I}_{x_0,\infty}(x) \cdot f_X(x)}{\mathbb{P}(X > x_0)} \, \mathrm{d}x = \int_{x_0}^\infty \frac{x \cdot f_X(x)}{\overline{F}(x_0)} \, \mathrm{d}x \tag{12.11}$$

To quantify $\mathbb{E}[T]$, we can split the process into two classes of events as follows:

1. **Type 1**: If $X \leq x_0$, which happens with probability $\mathbb{P}(X \leq x_0) = F(x_0)$, we ignore it.
2. **Type 2**: If $X > x_0$, which happens with probability $\mathbb{P}(X > x_0) = \overline{F}(x_0)$, we accept the first one that comes out way.

Thus, the arrival time of the first Type II event is distributed as $T \sim \textsc{exponential}(\lambda \cdot \overline{F}(x_0))$, which implies the following.

$$\mathbb{E}[T] = \frac{1}{\lambda \cdot \overline{F}(x_0)} \tag{12.12}$$

Thus, our total equation for $\mathbb{E}[R(x_0)]$ is as follows.

$$\mathbb{E}[R(x_0)] = \frac{1}{\overline{F}(x_0)} \cdot \left(\int_{x_0}^\infty x \cdot f(x) \, \mathrm{d}x - \frac{c}{\lambda} \right) \tag{12.13}$$

We now need to differentiate with respect to x_0 as follows.

$$\frac{d\mathbb{E}[R(x_0)]}{dx_0} = \frac{-x_0 \cdot f(x_0) \cdot \overline{F}(x_0) + \left(\int_{x_0}^\infty x \cdot f(x) \, \mathrm{d}x - \frac{c}{\lambda} \right) \cdot f(x_0)}{(\overline{F}(x_0))^2} \tag{12.14}$$

We can set the numerator equal to 0, and simplify as follows.

$$0 = -x_0 \cdot f(x_0) \cdot \overline{F}(x_0) + \left(\int_{x_0}^{\infty} x \cdot f(x) \, \mathrm{d}x - \frac{c}{\lambda} \right) \cdot f(x_0) \tag{12.15}$$

$$0 = -x_0 \cdot \overline{F}(x_0) + \int_{x_0}^{\infty} x \cdot f(x) \, \mathrm{d}x - \frac{c}{\lambda} \tag{12.16}$$

$$\frac{c}{\lambda} = -x_0 \cdot \int_{x_0}^{\infty} f(x) \, \mathrm{d}x + \int_{x_0}^{\infty} x \cdot f(x) \, \mathrm{d}x \tag{12.17}$$

$$\frac{c}{\lambda} = \int_{x_0}^{\infty} (x - x_0) \cdot f(x) \, \mathrm{d}x \equiv G(x_0) \tag{12.18}$$

One can observe that $G(x_0)$ is a decreasing function with respect to x_0, and $G(0) = \mathbb{E}[X]$. Therefore, our policy can be reduced to the following.

1. If $\mathbb{E}[X] < \frac{c}{\lambda}$, then we should accept the first offer.
2. If $\mathbb{E}[X] \geq \frac{c}{\lambda}$, then we utilize the x_0^* policy where x_0^* is the **unique** solution to Equation 12.18 that can be found numerically.

12.4 Conditional Arrival Times

Recall that in Lecture 12, we derived the distribution of the n-th arrival (waiting) time $S_n \sim \text{GAMMA}(n, \lambda)$, which is the time at which the n-th event occurs.

$$S_n = \sum_{i=1}^{n} T_i = S_{n-1} + T_n \tag{12.19}$$

In practice, we often do not know when people are exactly entering the system, but rather the total amount of people after some time has elapsed. It is of interest to characterize conditional distribution of $S_1, \cdots, S_n$ given $N_t = n$. It turns out that this distribution has a very nice result that is quite useful in applications.

We first expand via Bayes' rule as follows.

$$f_{S_1, \cdots, S_n \mid N_t}(s_1, \cdots, s_n \mid n) = \frac{f_{S_1, \cdots, S_n, N_t}(s_1, \cdots, s_n, n)}{\mathbb{P}(N_t = n)} \tag{12.20}$$

The distribution of $S_1, \ldots, S_n$ is complicated to deal with, so we would like to shift perspective to re-writing $S_1, \ldots, S_n$ in terms of $T_1, \ldots, T_n$. Using the recursive definition of S_n, we can do exactly this as follows.

$$T_n = S_{n+1} - S_n \tag{12.21}$$

Therefore, for any $0 < s_1 < \cdots < s_n < t$, we have the following equivalences. The last result is to enforce that there is no event that occurs in the time interval $(s_n, t]$ since we are given that $N_t = n$.

$$S_1 = s_1 \iff T_1 = s_1 \tag{12.22}$$

$$S_2 = s_2 \iff T_2 = s_2 - s_1 \tag{12.23}$$

$$\vdots \tag{12.24}$$

$$S_n = s_n \iff T_n = s_n - s_{n-1} \tag{12.25}$$

$$N_t = n \iff T_{n+1} > t - s_n \tag{12.26}$$

Thus, our conditional PDF now results in the following. We can now factorize this due to the i.i.d property of the interarrival times. From there, we can use a telescoping property on the product of densities to simplify greatly to our final result.

$$f_{S_1,\cdots,S_n \mid N_t}(s_1,\cdots,s_n \mid n) = \frac{f_{T_1,\cdots,T_n}(s_1,\cdots,s_n - s_{n-1}) \cdot \mathbb{P}(T_{n+1} > t - s_n)}{\mathbb{P}(N_t = n)} \tag{12.27}$$

$$= \frac{f_{T_1}(s_1)\cdots f_{T_n}(s_n - s_{n-1}) \cdot \mathbb{P}(T_{n+1} > t - s_n)}{\mathbb{P}(N_t = n)} \tag{12.28}$$

$$= \frac{\prod_{i=1}^{n} \lambda \cdot \exp(-\lambda \cdot (s_i - s_{i-1})) \cdot \exp(-\lambda \cdot (t - s_n))}{\exp(-\lambda t) \cdot \frac{(\lambda t)^n}{n!}} \tag{12.29}$$

$$= \frac{\exp\left(-\lambda \cdot \sum_{i=1}^{n}(s_i - s_{i-1})\right) \cdot \exp(\lambda s_n)}{\frac{(t)^n}{n!}} \tag{12.30}$$

$$= \frac{\exp(-\lambda s_n) \cdot \exp(\lambda s_n)}{\frac{(t)^n}{n!}} = \frac{n!}{t^n} \tag{12.31}$$

This answer is surprisingly nice given the complexity of the waiting time distribution, so it warrants an investigation via a concept known as an **order statistic**.

12.5 Order Statistics

A common concept in probability and statistics is **ordering** random variables.

Definition 12.2 (Order Statistics). The random variables $X_{(k)}$ represent the k-th smallest among $X_1, \cdots, X_n$, namely

$$X_{(1)} = \min\{X_1, \cdots, X_n\} \qquad X_{(n)} = \max\{X_1, \cdots, X_n\}$$

If $X_1, \cdots, X_n$ are i.i.d with PDF f, then it is easy to derive the joint PDF of $X_{(1)}, \cdots, X_{(n)}$. For a fixed $x_1 < x_2 < \cdots < x_n$, we can consider the probability that each of the $X_{(i)}$ are in a neighborhood $\mathcal{N}_\epsilon(\cdot)$ of the desired number. Here ϵ^n represents the differential volume of the region.

$$f_{X_{(1)},\cdots,X_{(n)}}(x_1,\cdots,x_n) \cdot \epsilon^n \approx \mathbb{P}(X_{(1)} \in \mathcal{N}_\epsilon(x_1),\cdots,X_{(n)} \in \mathcal{N}_\epsilon(x_n)) \tag{12.32}$$

It is clear that if we consider any permutation σ_k of the tuple $(x_1,\cdots,x_n)$, then this will satisfy the properties of having x_1 be the smallest, x_n the largest, and everything in between. Thus, it suffices to consider all permutations of the fixed tuple $(x_1,\cdots,x_n)$.

$$f_{X_{(1)},\cdots,X_{(n)}}(x_1,\cdots,x_n) \cdot \epsilon^n \approx \sum_{k=1}^{n!} \mathbb{P}(X_1 \in \mathcal{N}_\epsilon(x_{\sigma_k(1)}),\cdots,X_n \in \mathcal{N}_\epsilon(x_{\sigma_k(n)})) \tag{12.33}$$

With this insight, the joint density simplifies as follows.

$$f_{X_{(1)},\cdots,X_{(n)}}(x_1,\cdots,x_n) \cdot \epsilon^n \approx \sum_{k=1}^{n!} \prod_{i=1}^{n} \mathbb{P}(X_i \in \mathcal{N}_\epsilon(x_{\sigma_k(i)}) \tag{12.34}$$

$$\approx \sum_{k=1}^{n!} \prod_{i=1}^{n} \epsilon \cdot f_{X_i}(x_{\sigma_k(i)}) \tag{12.35}$$

$$= \sum_{k=1}^{n!} \epsilon^n \cdot \prod_{i=1}^{n} f_{X_i}(x_i) \tag{12.36}$$

$$= n!\epsilon^n \cdot \prod_{i=1}^{n} f_{X_i}(x_i) \tag{12.37}$$

Thus, in the $\epsilon \to 0$ limit, we obtain the following joint density on the order statistics.

$$f_{X_{(1)},\cdots,X_{(n)}}(x_1,\cdots,x_n) = n! \cdot \prod_{i=1}^{n} f_{X_i}(x_i) \tag{12.38}$$

The marginal density of each order statistic can then be written as follows.

$$f_{X_{(k)}}(x_k) = \sum_{j=k}^{n} \binom{n}{j} \cdot F(x)^j \cdot (1-F(x))^{n-j} \tag{12.39}$$

Aside: The intuition of this is that we need k of the random variables to be less than x_k, and $n-k-1$ of the random variables to be greater than x_k.

For $X_1,\cdots,X_n \sim \text{UNIFORM}(0,t)$ i.i.d, we then have that $f(x_i) = \frac{1}{t}$, which implies that

$$f_{X_{(1)},\cdots,X_{(n)}}(x_1,\cdots,x_n) = \frac{n!}{t^n} \tag{12.40}$$

With this new insight regarding order statistics, we have the following theorem regarding the conditional arrival times, which will be utilized next lecture.

Theorem 12.3 (Conditional Waiting Times). The conditional distribution of $S_1,\cdots,S_n$ given $N_t = n$ have the same distribution as the order statistics corresponding to n i.i.d random variables uniformly distributed on $(0,t)$. Equivalently, the **unordered** random variables $S_1,\cdots,S_n \sim \text{UNIFORM}(0,t)$ i.i.d given $N_t = n$.

13. Applications of Poisson Processes

Last lecture, we have established the following **important result**: if $\{N_t\}_{t\geq 0}$ is a Poisson process and $N_t = n$, then the arrival times $S_1, \cdots, S_n$ have the same distribution as the order statistics $U_{(1)}, \cdots, U_{(n)}$ corresponding to $U_1, \cdots, U_n \sim \text{UNIFORM}(0, t)$ i.i.d. Let us consider several applications of this result.

Agenda:
1. Insurance Claims
2. Multi-Type Poisson Processes
3. Health Care
4. Traffic Engineering

13.1 Insurance Claims

Suppose that the following simplifying assumptions on this model are made.

1. Insurance claims are made at times distributed according to a Poisson process $\{N_t\}_{t\geq 0}$ with rate λ.
2. The claim amounts are made i.i.d. with mean μ. This implies that if C_i represents the amount of the i-th claim, then $C(t)$ is the total cost up to time t.
$$C(t) = \sum_{i=1}^{N_t} C_i \tag{13.1}$$
3. If S_i represents the time of the i-th claim, then the "discounted price" $\tilde{C}_i$ of the i-th claim is as follows where α is the discounting rate that is determined by the company and $\chi(S_i) \in (0, 1)$ is a decreasing function.
$$\chi(S_i) = e^{-\alpha \cdot S_i} \tag{13.2}$$

Thus, the quantity of interest $D(t)$ is the total discounted price across all claims as follows.

$$D(t) = \sum_{i=1}^{N_t} \tilde{C}_i = \sum_{i=1}^{N_t} \chi(S_i) \cdot C_i = \sum_{i=1}^{N_t} e^{-\alpha \cdot S_i} \cdot C_i \tag{13.3}$$

We now seek to determine the expected value $\mathbb{E}[D(t)]$, which we can approach by conditioning on $N_t = n$.

$$\mathbb{E}[D(t)] = \mathbb{E}[\mathbb{E}[D(t) \mid N_t]]$$

To compute the conditional expectation, we substitute $D(t)$ to obtain the following. We use the fact that $U_1, \cdots, U_n \sim \text{UNIFORM}(0, t)$ i.i.d to go from Equation 13.5 to Equation 13.7.

$$\begin{aligned}
\mathbb{E}[D(t) \mid N_t = n] &= \mathbb{E}\left[\sum_{i=1}^{n} e^{-\alpha \cdot S_i} \cdot C_i \mid N_t = n\right] && (13.4)\\
&= \mathbb{E}\left[\sum_{i=1}^{n} e^{-\alpha \cdot U_{(i)}}\right] \cdot \mathbb{E}\left[C_i\right] && (13.5)\\
&= \mathbb{E}\left[\sum_{i=1}^{n} e^{-\alpha \cdot U_i}\right] \cdot \mu && (13.6)\\
&= \mathbb{E}\left[n \cdot e^{-\alpha \cdot U_1}\right] \cdot \mu && (13.7)\\
&= \frac{n\mu}{t} \cdot \int_0^t \exp(-\alpha \cdot x)\, \mathrm{d}x && (13.8)\\
&= \frac{n\mu}{\alpha t} \cdot \left(1 - e^{-\alpha t}\right) && (13.9)
\end{aligned}$$

Unfreezing the value of N_t and taking the expectation, we have the following.

$$\mathbb{E}[D(t)] = \mathbb{E}\left[\frac{N_t\mu}{\alpha t}\cdot\left(1-e^{-\alpha t}\right)\right] \tag{13.10}$$

$$= \frac{\lambda t\mu}{\alpha t}\cdot\left(1-e^{-\alpha t}\right) \tag{13.11}$$

$$= \boxed{\frac{\mu\lambda}{\alpha}\cdot\left(1-e^{-\alpha t}\right)} \tag{13.12}$$

13.2 Multi-Type Poisson Process

The last time we discussed the splitting of a Poisson process $\{N_t, t \geq 0\}$ into two independent processes. This construction can be directly generalized to any number of finite processes under the following assumptions:

1. $\{N_t\}_{t\geq 0}$ is a Poisson process with rate λ
2. Each event of type i has associated probability p_i for $i = 1, \cdots, m$

$$\sum_{i=1}^{m} p_i = 1 \tag{13.13}$$

This leads to the following decomposition.

1. $N_t^{(i)} = \#$ type i events in $[0,t]$ is a Poisson process with rate λp_i
2. $\{N_t^{(1)}, t \geq 0\}, \cdots, \{N_t^{(m)}, t \geq 0\}$ are independent

Suppose now, however, that probabilities p_i do depend on time the event occurs $p_i \equiv p_i(s)$. It is of interest to see if $\{N_t^{(i)}\}$ are still Poisson processes. This is clearly **not** true since $p_i(s)$ will most likely violate the stationary increments property, a clear example being if $p_i(s)$ is an increasing function of s.

Aside: For conservation of probability, we need that for each s

$$\sum_{i=1}^{m} p_i(s) = 1$$

However, not all hope is lost! Using the conditional waiting times derived in the previous lecture, one can show that the following is true.

1. $\{N_t^{(1)}\}_{t\geq 0}, \cdots, \{N_t^{(m)}\}_{t\geq 0}$ are independent random variables
2. $\{N_t^{(i)}\}_{t\geq 0}$ is a Poisson process with rate $\lambda_t^{(i)}$ where

$$\lambda_t^{(i)} = \lambda \cdot \int_0^t p_i(s)\, ds \tag{13.14}$$

This is known as an **inhomogenous** Poisson process, and we will study applications of this type of process in the following sections.

13.3 Estimating HIV Infections

The HIV virus causes AIDS, but it has a long incubation period: the time between the moment when an individual becomes infected and the moment when the symptoms of AIDS appears is large.

Thus, it is difficult to estimate the number of people that are infected at given time, but easy to estimate the number of people that have symptoms. To model this, we make the following simplifying assumptions:

1. Assume that individuals contact the HIV virus according to a Poisson process with unknown intensity λ.
2. Assume the time between being infected and having symptoms is modeled as a random variable $T \sim F_T(\cdot)$ i.i.d for all individuals.

Aside: While the Poisson assumption is reasonable, the constant λ is not. Furthermore, we assume access to F_T through empirical estimation.

Let $N_t^{\text{HIV}} = \#$ HIV positive individuals at time t, where $\{N_t^{\text{HIV}}, t \geq 0\} \sim \text{POISSON}(\lambda)$. We can decompose this into the sum of two processes as follows.

$$N_t^{\text{HIV}} = N_t^{\text{AIDS}} + N_t^{\text{no AIDS}} \tag{13.15}$$

We describe both of these processes as follows:

1. N_t^{AIDS} is an *observable* process that denotes the number of individuals with AIDS symptoms at time t
2. $N_t^{\text{no AIDS}}$ is an *non-observable* process that denotes the number of individuals with no symptoms at time t

The **goal** now is to estimate $N_t^{\text{no AIDS}}$. To do this, we can classify each event of contracting the virus at time s into two types:

1. **Type I**: A person has symptoms by time t w.p

$$\mathbb{P}(T \leq t - s) = F_T(t - s) \tag{13.16}$$

Aside: The notation $\overline{F}_T(\cdot) := 1 - F_T(\cdot)$ typically denotes the complementary CDF of a random variable.

2. **Type II**: A person does not have symptoms by time t w.p

$$\mathbb{P}(T > t - s) = 1 - F_T(t - s) = \overline{F}_T(t - s) \tag{13.17}$$

Thus, N_t^{AIDS} represents the number of Type 1 events by time t and $N_t^{\text{no AIDS}}$ represents the number of Type 2 events by time t.

$$\mathbb{E}[N_t^{\text{AIDS}}] = \lambda \cdot \int_0^t F_T(t - s)\, \mathrm{d}s = \lambda \cdot \int_0^t F_T(s)\, \mathrm{d}s \tag{13.18}$$

$$\mathbb{E}[N_t^{\text{no AIDS}}] = \lambda \cdot \int_0^t \overline{F}_T(t - s)\, \mathrm{d}s = \lambda \cdot \int_0^t \overline{F}_T(s)\, \mathrm{d}s \tag{13.19}$$

If λ was known, then we could directly infer $N_t^{\text{no AIDS}}$ by the quantity $\mathbb{E}[N_t^{\text{no AIDS}}]$. Since we do not have access to λ, we instead estimate based on the expected values.

$$\lambda = \frac{\mathbb{E}[N_t^{\text{AIDS}}]}{\int_0^t F_T(s)\, \mathrm{d}s} \implies \mathbb{E}[N_t^{\text{no AIDS}}] = \frac{\mathbb{E}[N_t^{\text{AIDS}}] \cdot \int_0^t \overline{F}_T(s)\, \mathrm{d}s}{\int_0^t F_T(s)\, \mathrm{d}s} \tag{13.20}$$

To get an estimate of $\mathbb{E}[N_t^{\text{no AIDS}}]$, we can estimate the expected value $\mathbb{E}[N_t^{\text{AIDS}}]$ by the actually observed number n_t^{AIDS} of patients with AIDS.

13.4 Minimizing Car Encounters

Cars enter a one-way highway of length d according to a Poisson process with rate λ, where N_t is the number of cars that have entered by time t. Thus, we have that $\{N_t\}_{t \geq 0} \sim \text{POISSON}(\lambda)$.

Each car travels at a constant speed, modeled as a random variable X with pdf $G(\cdot)$ i.i.d. When a faster car encounters a slower car, it passes with no time being lost.

Once our car enters the highway at time $t = \tau$, we can choose the speed x of the car. We now seek to determine the optimal speed x that minimizes the expected number of encounters with other cars.

To do this, we can classify each car entering the highway as one of two types. Note that we are interested in the first process as we wish to characterize the number of cars we pass.

- **Type 1**: We encounter the car, noted by $\{N_t^{(1)}\}_{t\geq 0}$.
- **Type 2**: We do not encounter the car, noted by $\{N_t^{(2)}\}_{t\geq 0}$.

If we enter the highway at time τ, then we will exist at time $\tau + \frac{d}{x}$. Any cars ahead of us at time $\tau + \frac{d}{x}$ will not be be seen, which means it suffices to minimize the following quantity.

$$\mathbb{E}\left[N^{(1)}_{\tau+\frac{d}{x}}\right] \tag{13.21}$$

Since $\{N_t^{(1)}\}_{t\geq 0}$ is an inhomogenous Poisson process, we can simplify as follows where $p_1(s)$ is the probability we encountered a car that entered at time t.

$$\mathbb{E}\left[N^{(1)}_{\tau+\frac{d}{x}}\right] = \lambda^{(1)}_{\tau+\frac{d}{x}} = \lambda \cdot \int_0^{\tau+\frac{d}{x}} p_1(t)\,\mathrm{d}t \tag{13.22}$$

The two cases here are that

1. a car entered before our car and we pass it
2. a car entered after our car and passes us.

For **1**), this case is equivalent to computing the following probability where $t < \tau$ and T is the travel time of the car. Using the fact that $T := d/X$, we can simplify as follows.

$$p_1(t) = \mathbb{P}\left(t + T > \tau + \frac{d}{x}\right) = \mathbb{P}\left(X < \frac{d}{\tau + \frac{d}{x} - t}\right) = G\left(\frac{d}{\tau + \frac{d}{x} - t}\right) \tag{13.23}$$

For **2**), this case is equivalent to computing the following probability for $t > \tau$ and T the travel time of the car using nearly identical computation.

$$p_1(t) = \mathbb{P}\left(t + T < \tau + \frac{d}{x}\right) = \mathbb{P}\left(X > \frac{d}{\tau + \frac{d}{x} - t}\right) = \overline{G}\left(\frac{d}{\tau + \frac{d}{x} - t}\right) \tag{13.24}$$

Thus, our expectation reduces to the following upon splitting the interval of integration.

$$\mathbb{E}\left[N^{(1)}_{\tau+\frac{d}{x}}\right] = \lambda \cdot \int_0^{\tau} p_1(t)\,\mathrm{d}t + \lambda \cdot \int_\tau^{\tau+\frac{d}{x}} p_1(t)\,\mathrm{d}t \tag{13.25}$$

$$= \lambda \cdot \int_0^{\tau} G\left(\frac{d}{\tau + \frac{d}{x} - t}\right)\mathrm{d}t + \lambda \cdot \int_\tau^{\tau+\frac{d}{x}} \overline{G}\left(\frac{d}{\tau + \frac{d}{x} - t}\right)\mathrm{d}t \tag{13.26}$$

$$= \lambda \cdot \int_{\frac{d}{x}}^{\tau+\frac{d}{x}} G\left(\frac{d}{s}\right)\mathrm{d}s + \lambda \cdot \int_0^{\frac{d}{x}} \overline{G}\left(\frac{d}{s}\right)\mathrm{d}s \qquad \left(s = \tau - t + \frac{d}{x}\right) \tag{13.27}$$

To make our lives a tad easier, we will consider $y := d/x$. This yields the following.

$$\mathbb{E}\left[N^{(1)}_{\tau+y}\right] = \lambda \cdot \int_y^{\tau+y} G\left(\frac{d}{s}\right) \, \mathrm{d}s + \lambda \cdot \int_0^y \overline{G}\left(\frac{d}{s}\right) \, \mathrm{d}s \tag{13.28}$$

We can now use Fundamental Theorem of Calculus to take the derivative of this quantity with respect to y in order to yield the following.

$$\frac{\mathrm{d}\mathbb{E}\left[N^{(1)}_{\tau+y}\right]}{\mathrm{d}y} = \lambda \cdot \left(G\left(\frac{d}{\tau+y}\right) - G\left(\frac{d}{y}\right) + \overline{G}\left(\frac{d}{y}\right)\right) = 0 \tag{13.29}$$

One can then solve this equation for y to obtain the optimal travel time y^*, which will in turn yield the optimal speed x^*. More interestingly, the long-time behavior of this system simplifies quite nicely if we assume that $\tau \to \infty$. This turns off the first term in our derivative, which yields the following.

$$0 = \lambda \cdot \left(-G\left(\frac{d}{y}\right) + \overline{G}\left(\frac{d}{y}\right)\right) \tag{13.30}$$

$$= \lambda \cdot \left(-G\left(\frac{d}{y}\right) + 1 - G\left(\frac{d}{y}\right)\right) \tag{13.31}$$

$$= \lambda \cdot \left(1 - 2 \cdot G\left(\frac{d}{y}\right)\right) \tag{13.32}$$

This yields the following.

$$G\left(\frac{d}{y}\right) = G(x) = \frac{1}{2} \tag{13.33}$$

Thus, as we expect, the optimal speed is the **median** of G.

14. Brownian Motion

The goal of this lecture is to introduce Brownian motion, which is a another widely-used stochastic process whose applications range from quantum mechanics to pricing stock options, modeling dynamics of particles in a fluid to internet traffic. We will discuss only the basic properties of Brownian motion since its deeper understanding requires Itô calculus (the calculus for stochastic processes), which is out of the scope of the course (covered in CMS 117/SS 214).

Agenda:
1. Continuum Limit of Random Walk
2. Properties of Brownian Motion
3. Hitting Times and Maximum Values
4. Variants of Brownian Motion

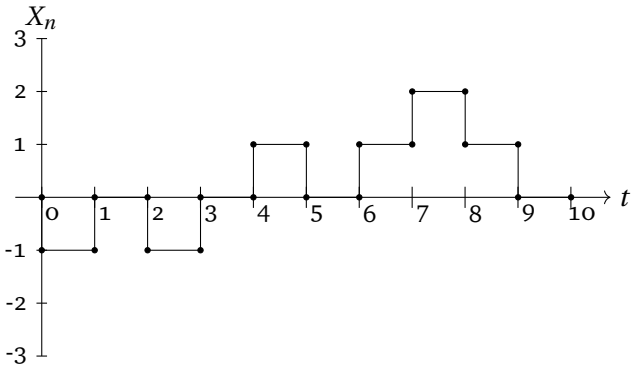
14.1 Continuum Limit of Random Walk

Consider the following discrete-time stochastic process $\{X_n\}_{n\in\mathbb{N}}$, which is known as a symmetric random walk.

Aside: We say that the random variable $Z \sim$ Rademacher, which is a distribution often used for modeling random walks.

$$\begin{cases} X_0 = 0 \\ X_{n+1} = X_n + Z \quad \text{where } Z = \begin{cases} -1 & \text{w.p } \frac{1}{2} \\ 1 & \text{w.p } \frac{1}{2} \end{cases} \end{cases} \tag{14.1}$$

Below is a realization of $\{X_n\}_{n\in[0,10]}$, which looks as you expect from incrementing up or down by 1 at every time step.

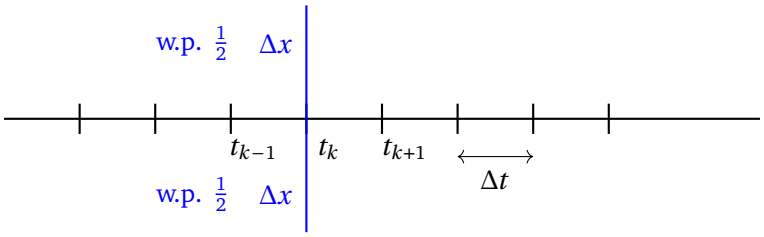


In nature, the interactions we are interested in modeling are happening in fractions of picoseconds. Thus, it is of interest to take a limit in the both space and time to recover something that could model the continuum and recover the Brownian motion, which can be done by taking smaller steps in both space (Δx) and time (Δt).

Aside: The function $\lfloor x \rfloor$ is the largest integer $\leq x$. For some examples, $\lfloor 5.5 \rfloor = 5$ and $\lfloor 10.9 \rfloor = 10$.

$$\begin{cases} X_0 = 0 \\ X_t = \Delta x \cdot \sum_{i=1}^{\lfloor t/\Delta t \rfloor} Z_i \qquad Z_i \sim \text{Rademacher i.i.d} \end{cases} \tag{14.2}$$

Consider the visual below to see a zoomed-in version of the motion, in which it is equally probable to move up or down by Δx per time step Δt.



We now need to consider how "fast" we take the limit of $\Delta t \to 0$ and $\Delta x \to 0$. Computing the expectation and variance of X_t, we have the following.

$$\mathbb{E}[X_t] = \Delta x \cdot \sum_{i=1}^{\lfloor t/\Delta t \rfloor} \mathbb{E}[Z_i] = 0 \tag{14.3}$$

$$\mathbb{V}[X_t] = (\Delta x)^2 \cdot \sum_{i=1}^{\lfloor t/\Delta t \rfloor} \mathbb{V}[Z_i] = (\Delta x)^2 \cdot \sum_{i=1}^{\lfloor t/\Delta t \rfloor} 1 = (\Delta x)^2 \cdot \lfloor t/\Delta t \rfloor \tag{14.4}$$

To get a non-trivial limit, we will simply let $\Delta x = \sigma \cdot \sqrt{\Delta t}$ for $\sigma > 0$, which yields the following as $\Delta t \to 0$.

$$\mathbb{V}[X_t] = (\sigma \cdot \sqrt{\Delta t})^2 \cdot \lfloor t/\Delta t \rfloor = \sigma^2 \cdot t \tag{14.5}$$

We can now ask what happens in the limiting process $\{X_t\}_{t \geq 0}$, for which we can immediately use CLT to say the following.

1. $X_0 = 0$
2. $X_t \sim \text{NORMAL}(0, \sigma^2 t)$

We now prove the desired properties of a stochastic process we have discussed up to this point, which are independence and stationarity of increments. To show the former, we can consider the difference.

$$X_{t_2} - X_{t_1} = \Delta x \cdot \sum_{i=1}^{\lfloor t_2/\Delta t \rfloor} Z_i - \Delta x \cdot \sum_{i=1}^{\lfloor t_1/\Delta t \rfloor} Z_i = \Delta x \cdot \sum_{i=\lfloor t_1/\Delta t \rfloor + 1}^{\lfloor t_2/\Delta t \rfloor} Z_i \tag{14.6}$$

Since the $\{Z_i\}$ are independent, the increments will also be independent. As for the latter, we consider the CLT on $X_{t_2} - X_{t_1}$ with $n = \lfloor t_2/\Delta t \rfloor - \lfloor t_1/\Delta t \rfloor \approx (t_2 - t_1)/\Delta t$ terms.

$$X_{t_2} - X_{t_1} \sim \text{NORMAL}(0, \sigma^2(t_2 - t_1)) \tag{14.7}$$

This motivates the following formal definition of this limiting process.

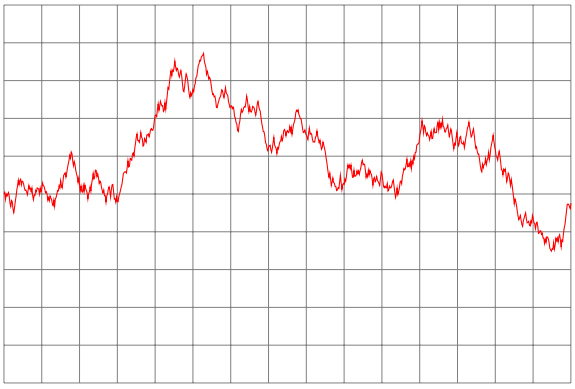
Definition 14.1 (Brownian Motion). A stochastic process $\{X_t\}_{t \geq 0}$ is said to be a Brownian motion process with variance σ^2 if it satisfies the following properties:

1. $X_0 = 0$
2. $\{X_t\}_{t \geq 0}$ has independent and stationary increments
3. $X_t \sim \text{NORMAL}(0, \sigma^2 t)$ for every $t \geq 0$

For some context, here is the history of how Brownian motion was developed.

- **1827**: Brownian motion was discovered by biologist Robert Brown. He was studying pollen particles floating in water and observed the jittery motion of the particles.
- **1905**: Albert Einstein developed the first probability model that could explain Brownian motion.
- **1918**: Norbert Weiner gave the above definition and developed the theory of Brownian motion, which is also called the Weiner process.

Henceforth, we will use the typical notation for a Brownian motion with either $\{B_t\}_{t\geq 0}$ or $\{W_t\}_{t\geq 0}$. Here is an example trajectory of Brownian motion, which may look familiar to those who have worked with stocks.



While the definition of Brownian motion seems somewhat ordinary, its properties are anything but! We will discuss some of them now.

Aside: These proofs again involve stochastic analysis, so we will give heuristic arguments as to why they are true.

Theorem 14.2 (Continuity of BM). The function $t \mapsto B_t$ is continuous w.p. 1.

Proof. **Intuitively**, continuity of $f(t)$ means that the change of $|f(t+\Delta t) - f(t)|$ is small if Δt is small. With the definition, we can compute the distribution as follows.

$$B_{t+\Delta t} - B_t \sim \text{NORMAL}(0, \sigma^2 \Delta t) \tag{14.8}$$

This implies that $\mathbb{E}[(B_{t+\Delta t} - B_t)^2] = \sigma^2 \Delta t$, which in turn gives us the expected deviation of the two values.

$$|B_{t+\Delta t} - B_t| \approx \sigma \cdot \sqrt{\Delta t} \tag{14.9}$$

Since $\sigma \cdot \sqrt{\Delta t} \to 0$ as $\Delta t \to 0$, this proves continuity. ∎

Theorem 14.3 (Differentiability of BM). The function $t \mapsto B_t$ is nowhere differentiable w.p. 1.

Proof. **Intuitively**, we can write the limit definition of the derivative as follows.

$$\frac{\mathrm{d}B_t}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{B_{t+\Delta t} - B_t}{\Delta t} \sim \lim_{\Delta t \to 0} \frac{\sigma \cdot \sqrt{\Delta t}}{\Delta t} = \infty \tag{14.10}$$

∎

If $\sigma = 1$, then $\{B_t\}_{t\geq 0}$ is called the **standard** Brownian motion. It is clear that any Brownian motion $\{B'_t\}_{t\geq 0}$ with variance σ^2 can be scaled to a standard Brownian motion by letting $B_t \equiv B'_t/\sigma$.

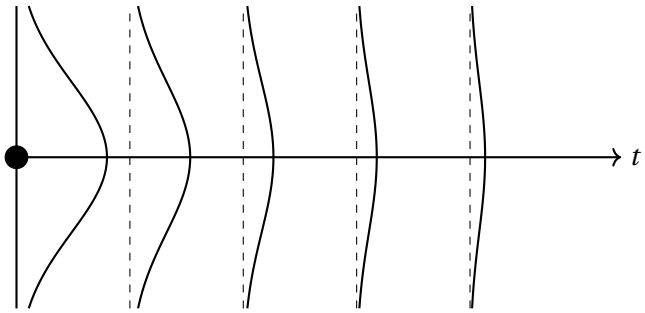
14.2 Properties of Brownian Motion

We will now discuss the main properties of the standard Brownian motion as follows.

Property 14.4

$$B_t \sim \text{NORMAL}(0,t) \implies f_{B_t}(x) = \frac{1}{\sqrt{2\pi t}} \cdot \exp\left(\frac{-x^2}{2t}\right)$$

Consider the inflation of variance as time passes on through the following visual.



Property 14.5

$$\text{Cov}(B_t, B_s) = \min\{s,t\}$$

Proof. This proof will use independent and stationary increments to simplify the expression. Writing out the covariance, we have the following by definition of Brownian motion.

$$\text{Cov}(B_s, B_t) = \mathbb{E}[B_s \cdot B_t] - \mathbb{E}[B_s] \cdot \mathbb{E}[B_t] = \mathbb{E}[B_s B_t] \tag{14.11}$$

The general strategy to split the timeline into two intervals, namely $[0, \min\{s,t\}]$ and $[\min\{s,t\}, \max\{s,t\}]$. From this, we obtain independent increments and can use a lot of nice properties. Thus, we can do the adding 0 trick as follows.

$$\begin{aligned}
\mathbb{E}[B_{\min\{s,t\}} B_{\max\{s,t\}}] &= \mathbb{E}[B_{\min\{s,t\}} \cdot (B_{\max\{s,t\}} - B_{\min\{s,t\}} + B_{\min\{s,t\}})] && (14.12)\\
&= \mathbb{E}[B_{\min\{s,t\}} \cdot (B_{\max\{s,t\}} - B_{\min\{s,t\}})] + \mathbb{E}[B^2_{\min\{s,t\}}] && (14.13)\\
&= \mathbb{E}[B_{\min\{s,t\}}] \cdot \mathbb{E}[(B_{\max\{s,t\}} - B_{\min\{s,t\}})] + \mathbb{E}[B^2_{\min\{s,t\}}] && (14.14)\\
&= \mathbb{E}[B^2_{\min\{s,t\}}] && (14.15)\\
&= \mathbb{V}[B_{\min\{s,t\}}] + \mathbb{E}[B_{\min\{s,t\}}]^2 = \min\{s,t\} && (14.16)
\end{aligned}$$

If $\sigma \neq 1$, then one can show that the covariance is $\text{Cov}(B_t, B_s) = \sigma^2 \cdot \min\{s,t\}$. ■

Property 14.6 For $t_1 < \cdots < t_n$, the joint PDF of $B_{t_1}, \cdots, B_{t_n}$ is

$$f_{B_{t_1},\cdots,B_{t_n}}(x_1,\cdots,x_n) = \frac{\exp\left(\frac{-1}{2} \cdot \left(\frac{x_1^2}{t_1} + \frac{(x_2-x_1)^2}{t_2-t_1} + \cdots + \frac{(x_n-x_{n-1})^2}{t_n-t_{n-1}}\right)\right)}{(\sqrt{2\pi})^n \cdot \sqrt{t_1(t_2-t_1)\cdots(t_n-t_{n-1})}}$$

Aside: Notice the similarity to the Poisson process covariance result.

$$\text{Cov}(N_s, N_t) = \lambda \cdot \min\{s,t\}$$

Proof. We can use independent increments to consider the translated distribution as follows.

$$B_{t_1} = x_1 \iff B_{t_1} = x_1 \tag{14.17}$$
$$B_{t_2} = x_2 \iff B_{t_2} - B_{t_1} = x_2 - x_1 \tag{14.18}$$
$$\vdots \tag{14.19}$$
$$B_{t_n} = x_n \iff B_{t_n} - B_{t_{n-1}} = x_n - x_{n-1} \tag{14.20}$$

Thus, the joint PDF can be translated to the increments above as follows where $x_0, t_0 = 0$.

$$f_{B_{t_1},\cdots,B_{t_n}}(x_1,\cdots,x_n) = f_{B_{t_1},\cdots,B_{t_n}-B_{t_{n-1}}}(x_1,\cdots,x_n - x_{n-1}) \tag{14.21}$$
$$= f_{B_{t_1}}(x_1)\cdots f_{B_{t_n}-B_{t_{n-1}}}(x_n - x_{n-1}) \tag{14.22}$$
$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(t_i - t_{i-1})}} \cdot \exp\left(\frac{-(x_i - x_{i-1})^2}{2(t_i - t_{i-1})}\right) \tag{14.23}$$
$$= \frac{1}{(\sqrt{2\pi})^n} \cdot \left(\prod_{i=1}^{n} \frac{1}{\sqrt{t_i - t_{i-1}}}\right) \cdot \exp\left(\sum_{i=1}^{n} \frac{-(x_i - x_{i-1})^2}{2(t_i - t_{i-1})}\right) \tag{14.24}$$
$$= \frac{\exp\left(\frac{-1}{2} \cdot \left(\frac{x_1^2}{t_1} + \frac{(x_2-x_1)^2}{t_2-t_1} + \cdots + \frac{(x_n-x_{n-1})^2}{t_n-t_{n-1}}\right)\right)}{(\sqrt{2\pi})^n \cdot \sqrt{t_1(t_2-t_1)\cdots(t_n-t_{n-1})}} \tag{14.25}$$

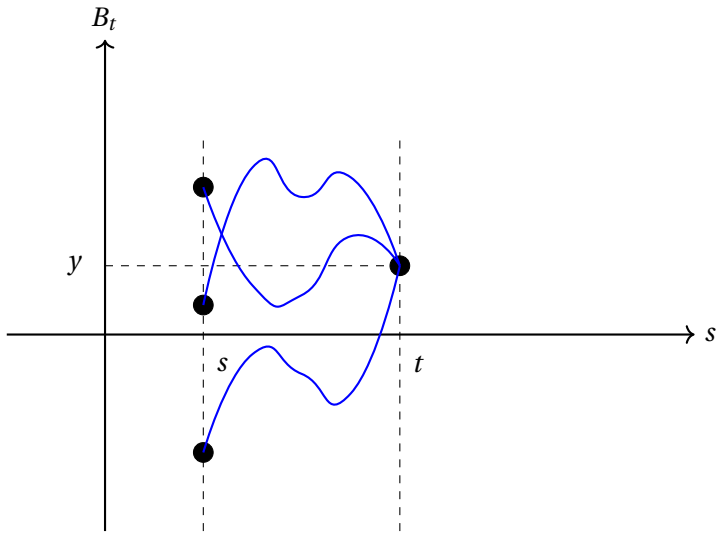
■

Property 14.7 The conditional distribution of B_s given by $B_t = y$, $s < t$

$$B_s \mid B_t = y \sim \text{NORMAL}\left(y \cdot \frac{s}{t}, (t-s)\cdot\frac{s}{t}\right)$$

Aside: A typical technique in conditioned stochastic processes is to observe what happens when one increases the gap between s and t.

Proof. We are interested in trying to predict previous information in the stochastic processes using future values, which is pictured below.



Using Baye's rule, we can expand the following in terms of the density.

$$f_{B_s \mid B_t}(x \mid y) = \frac{f_{B_t, B_s}(x, y)}{f_{B_t}(y)} \tag{14.26}$$

$$= \frac{\frac{1}{2\pi \cdot \sqrt{s(t-s)}} \cdot \exp\left(\frac{-1}{2} \cdot \left(\frac{x^2}{s} + \frac{(y-x)^2}{t-s}\right)\right)}{\frac{1}{\sqrt{2\pi t}} \cdot \exp\left(\frac{-y^2}{2t}\right)} \tag{14.27}$$

$$= \frac{1}{\sqrt{2\pi \cdot \frac{s}{t}(t-s)}} \cdot \exp\left(\frac{-1}{2} \cdot \left(\frac{x^2}{s} + \frac{(y-x)^2}{t-s} - \frac{y^2}{t}\right)\right) \tag{14.28}$$

$$= \frac{1}{\sqrt{2\pi \cdot \frac{s}{t}(t-s)}} \cdot \exp\left(\frac{-1}{2 \cdot \frac{s}{t}(t-s)} \cdot \left(x^2 - \frac{2syx}{t} + \frac{s^2 y^2}{t^2}\right)\right) \tag{14.29}$$

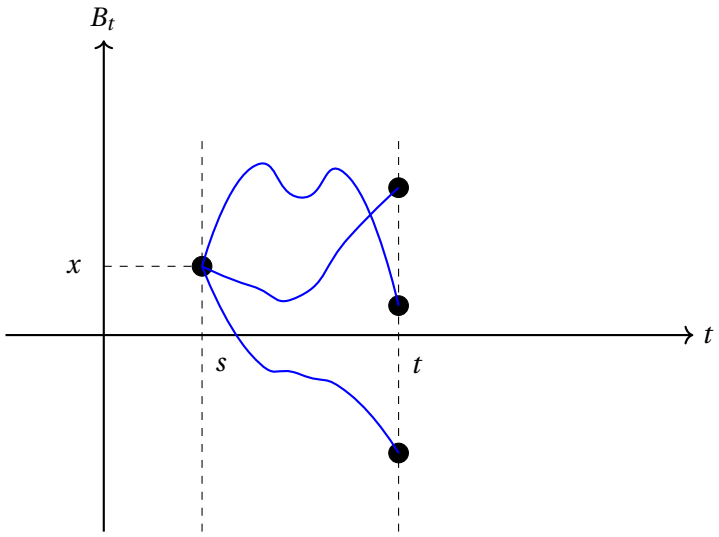
$$= \frac{1}{\sqrt{2\pi \cdot \frac{s}{t}(t-s)}} \cdot \exp\left(\frac{-1}{2 \cdot \frac{s}{t}(t-s)} \cdot \left(x - \frac{sy}{t}\right)^2\right) \tag{14.30}$$

■

Property 14.8 The conditional distribution of B_t given by $B_s = x$, $s < t$

$$B_t \mid B_s = x \sim \text{NORMAL}\,(x, t-s)$$

Proof. We are interested in trying to predict future information in the stochastic processes using past values, which is pictured below.



An intuitive proof of this is noting that $B_s \perp\!\!\!\perp B_t - B_s \sim \text{NORMAL}(0, t-s)$. Thus, one can deduce the following.

$$B_t - B_s \mid B_s = x \sim \text{NORMAL}(0, t-s) \tag{14.31}$$

$$B_t - x \mid B_s = x \sim \text{NORMAL}(0, t-s) \tag{14.32}$$

$$B_t \mid B_s = x \sim \text{NORMAL}(x, t-s) \tag{14.33}$$

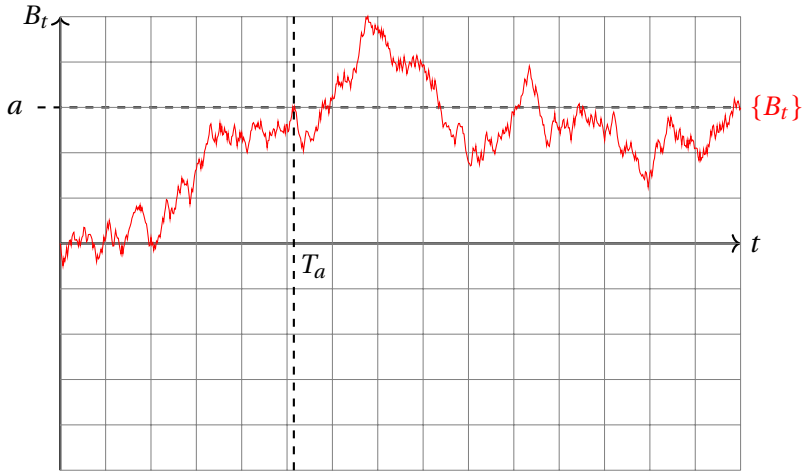
■

Aside: One can also prove this explicitly using Baye's rule as we did for the previous property.

This gives us tools to inspect how past and future information influence the probabilities of the present.

14.3 Hitting Times and Maximum Values

One random variable that often appears in applications and proofs in stochastic analysis is hitting time. Let $a > 0$, and let T_a be the random variable that represents the first time the Brownian motion $\{B_t\}$ hits a.



To obtain the distribution of T_a, we use the total law of probability on the event of $B_t \geq a$.

$$\mathbb{P}(B_t \geq a) = \mathbb{P}(B_t \geq a \mid T_a \leq t) \cdot \mathbb{P}(T_a \leq t) + \mathbb{P}(B_t \geq a \mid T_a > t) \cdot \mathbb{P}(T_a > t) \quad (14.34)$$

We see that $\mathbb{P}(B_t \geq a \mid T_a > t) = 0$, since $\{B_t\}$ cannot hit a by time t if the hitting time is after t. Thus, this simplifies as follows.

$$\mathbb{P}(B_t \geq a) = \mathbb{P}(B_t \geq a \mid T_a \leq t) \cdot \mathbb{P}(T_a \leq t) \quad (14.35)$$

We will now formally prove the value of the other conditional probability. Informally, one can say that $B_s = a$ for some $s \in [0, t]$, and is equally likely to be above or below this value at time t.

Lemma 14.9

$$\mathbb{P}(B_t \geq a \mid T_a \leq t) = \frac{1}{2}$$

Proof. We condition on the value of $T_a = s$ and integrate across this interval $s \in [0, t]$.

$$\mathbb{P}(B_t \geq a \mid T_a \leq t) = \int_0^t \mathbb{P}(B_t \geq a \mid T_a \leq t, T_a = s) \cdot f_{T_a \mid T_a \leq t}(s)\, \mathrm{d}s \quad (14.36)$$

We can re-express the conditional probability as follows.

$$\begin{aligned}
\mathbb{P}(B_t \geq a \mid T_a \leq t, T_a = s) &= \mathbb{P}(B_t \geq a \mid B_s = a, B_q < a \ \forall q < s) && (14.37)\\
&= \mathbb{P}(B_t - B_s \geq 0 \mid B_q < a \ \forall q < s) && (14.38)\\
&= \mathbb{P}(B_t - B_s \geq 0) && (14.39)\\
&= \frac{1}{2} \qquad (B_t - B_s \sim \text{NORMAL}(0, t - s)) && (14.40)\\
\mathbb{P}(B_t \geq a \mid T_a \leq t) &= \int_0^t \frac{1}{2} \cdot f_{T_a \mid T_a \leq t}(s)\, \mathrm{d}s = \frac{1}{2} && (14.41)
\end{aligned}$$

■

Thus, the original probability becomes the following.

$$\mathbb{P}(B_t \geq a) = \frac{1}{2} \cdot \mathbb{P}(T_a \leq t) \implies \mathbb{P}(T_a \leq t) = 2 \cdot \mathbb{P}(B_t \geq a) \tag{14.42}$$

Using the density of B_t, this probability becomes the following.

$$\begin{aligned}
\mathbb{P}(T_a \leq t) &= 2 \cdot \int_a^\infty \frac{1}{\sqrt{2\pi t}} \cdot \exp\left(\frac{-x^2}{2t}\right) \mathrm{d}x && && (14.43)\\
&= 2 \cdot \int_{a/\sqrt{t}}^\infty \frac{1}{\sqrt{2\pi}} \cdot \exp\left(\frac{-y^2}{2}\right) \mathrm{d}y && (x = y\sqrt{t}) && (14.44)\\
&= 2 \cdot \mathbb{P}\left(Z > \frac{a}{\sqrt{t}}\right) && (Z \sim \text{NORMAL}(0,1)) && (14.45)\\
&= 2 \cdot \left(1 - \Phi\left(\frac{a}{\sqrt{t}}\right)\right) && && (14.46)
\end{aligned}$$

It is clear that $T_a \sim T_{-a}$, which means that for $a \in \mathbb{R}$ with $a \neq 0$ and $t > 0$, the following is the CDF for the hitting time.

Theorem 14.10 (Hitting Time Distribution).

$$F_{T_a}(t) = \mathbb{P}(T_a \leq t) = 2 \cdot \left(1 - \Phi\left(\frac{|a|}{\sqrt{t}}\right)\right)$$

We can do a limiting behavior analysis of this CDF.

1. In the limit $t \to \infty$, we will almost surely at some point hit the fixed value of a. This also matches the formula we have in the limit.

$$F_{T_a}(\infty) = 2 \cdot \left(1 - \Phi\left(\frac{|a|}{\infty}\right)\right) = 2 \cdot \left(1 - \frac{1}{2}\right) = 1 \tag{14.47}$$

2. In the limit $a \to \infty$, we will almost never see the value of a for a fixed time t. This also matches the formula we have in the limit.

$$F_{T_\infty}(t) = 2 \cdot \left(1 - \Phi\left(\frac{|\infty|}{\sqrt{t}}\right)\right) = 2 \cdot (1 - 1) = 0 \tag{14.48}$$

Another quantity of interest is the **maximum** value of a Brownian motion attains in the interval $[0, t]$, formulated as follows.

$$M_t = \max_{s \in [0,t]} B_s \in [0, \infty) \tag{14.49}$$

We can use the hitting time definition to exactly quantify this CDF.

Theorem 14.11 (Maximum Value Distribution).

$$\mathbb{P}(M_t \geq m) = \mathbb{P}(T_m \leq t) = 2 \cdot \left(1 - \Phi\left(\frac{m}{\sqrt{t}}\right)\right)$$

14.4 Variations of Brownian Motion

If we want to have a more expressive class of motions, we can add a *drift* term μ to the motion through the following definition.

Definition 14.12 (Brownian Motion with Drift). A stochastic process $\{X_t, t \geq 0\}$ is said to be a Brownian motion process with drift if it satisfies the following properties:

1. $X_0 = 0$
2. $\{X_t\}$ has stationary and independent increments
3. $X_t \sim \text{NORMAL}(\mu t, \sigma^2 t)$

One can also scale a standard Brownian motion $\{B_t, t \geq 0\}$ to one with drift as follows.

$$X_t = \mu t + \sigma B_t \tag{14.50}$$

Another class of motions of particular interest is *geometric* Brownian motion, which can be used to model stock prices over time.

Definition 14.13 (Geometric Brownian Motion). A stochastic process $\{Y_t, t \geq 0\}$ is said to be a geometric Brownian motion process if it satisfies the following properties:

1. $Y_0 = 1$
2. $\{Y_t\}$ has stationary and independent increments
3. $Y_t \sim \exp(X_t)$ where X_t is a Brownian motion with drift

15. General Stochastic Processes

In the last four lectures, we discussed the main properties and some applications of Poisson processes in detail, as well as some basic properties of Brownian motion. For the rest of the course, we will focus on more **general** properties of stochastic processes.

Agenda:
1. Mean Functions
2. Covariance and Correlation Functions
3. Stationarity

15.1 Mean Functions

To set the stage, let $\{X_t\}_{t\in\mathsf{T}}$ be a stochastic process, or a collection of random variables where T is some index set of times.

Definition 15.1 (Mean Function). The **mean** function of $\{X_t\}_{t\in\mathsf{T}}$ is

$$m_X(t) = \mathbb{E}[X_t]$$

Intuitively, the mean function $m_X(t)$ reflects the *average* behavior of the process.

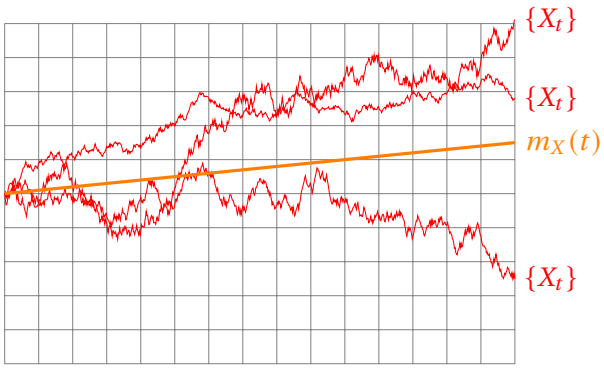
Example 15.2 Consider some examples for mean functions of processes.

1. **Poisson Process:** For $\{X_t\}_{t\geq 0}$ such that $X_t \sim \text{POISSON}(\lambda t)$,
$$m_X(t) = \lambda t \tag{15.1}$$
2. **Brownian Motion:** For $\{X_t\}_{t\geq 0}$ such that $X_t \sim \text{NORMAL}(\mu t, \sigma^2 t)$,
$$m_X(t) = \mu t \tag{15.2}$$
3. **Random Phase:** For $\{X_t\}_{t\in\mathbb{R}}$ such that $X_t = \cos(2\pi f t + \Theta)$, noting that $\Theta \sim \text{UNIFORM}[-\pi, \pi]$ and frequency f,
$$m_X(t) = \mathbb{E}[X_t] = \int_{-\pi}^{\pi} \cos(2\pi f t + \Theta) \cdot \frac{1}{2\pi}\, d\Theta = 0 \tag{15.3}$$

▪

Aside: The random phase process is often used as a model for the carrier signal in radio communications.

It is clear that even though the sample trajectories of a stochastic process are random, the mean function will be deterministic by the parameters of the process. The visual of different trajectories against the mean function is pictured below.



15.2 Covariance and Correlation Functions

We can now define the covariance function for a stochastic process as well.

Definition 15.3 (Covariance Function). The **covariance** function of $\{X_t\}_{t\in\mathsf{T}}$ is

$$C_X(t_1,t_2) = \mathbb{E}[(X_{t_1} - m_X(t_1)) \cdot (X_{t_2} - m_X(t_2))]$$

In Lecture 12, we computed $C_X(t_1,t_2) = \lambda \cdot \min\{t_1,t_2\}$ for the Poison process. Equivalently, we can express the covariance function as follows.

$$C_X(t_1,t_2) = \mathbb{E}[X_{t_1}X_{t_2}] - m_X(t_1)\cdot m_X(t_2) \tag{15.4}$$

In the study of stochastic processes, it is often assumed that $m_X(t) = 0$, since this will simplify these computations. This is not a particularly strong assumption since we can always

1. define a transformed process $\tilde{X}_t = X_t - m_X(t)$ with $m_{\tilde{X}}(t) = 0$
2. study $\tilde{X}_t$ and properties
3. use those properties to then deduce properties X_t

This motivates the following definition.

Definition 15.4 (Correlation Function). The **correlation** function of $\{X_t\}_{t\in\mathsf{T}}$ is

$$R_X(t_1,t_2) = \mathbb{E}[X_{t_1}X_{t_2}]$$

Aside: We will later show that the Fourier transform of the correlation function will reflect how "wiggly" a process is.

The correlation and covariance function are related as follows.

$$C_X(t_1,t_2) = R_X(t_1,t_2) - m_X(t_1)\cdot m_X(t_2) \tag{15.5}$$

Example 15.5 We consider the same processes as in the previous example.

1. **Poisson Process:** For $\{X_t\}_{t\geq 0}$ such that $X_t \sim \text{POISSON}(\lambda t)$,

$$R_X(t_1,t_2) = \lambda \min\{t_1,t_2\} + \lambda^2 t_1 t_2 \tag{15.6}$$

2. **Brownian Motion:** For $\{X_t\}_{t\geq 0}$ such that $X_t \sim \text{NORMAL}(\mu t, \sigma^2 t)$,

$$R_X(t_1,t_2) = \sigma^2 \min\{t_1,t_2\} + \mu^2 t_1 t_2 \tag{15.7}$$

3. **Random Phase:** For $\{X_t\}_{t\in\mathbb{R}}$ such that $X_t = \cos(2\pi f t + \Theta)$, noting that $\Theta \sim \text{UNIFORM}[-\pi,\pi]$ and frequency f,

$$R_X(t_1,t_2) = \mathbb{E}[\cos(2\pi f t_1 + \Theta)\cdot\cos(2\pi f t_2 + \Theta)] \tag{15.8}$$

   Using the product-to-sum identity, we derive the following.

$$R_X(t_1,t_2) = \frac{\cos(2\pi f(t_1 - t_2))}{2} \tag{15.9}$$

4. **Independent Sum**: For $\{X_n\}_{n\in\mathbb{N}}$ such that X_n is a sum of independent normal random variables.

$$X_n = \sum_{i=1}^{n} Z_i \qquad \{Z_i\}_{i=1}^{n} \sim \text{NORMAL}(0,\sigma^2) \text{ i.i.d} \tag{15.10}$$

Aside: This process is of interest as it is the discretization of the Brownian motion.

It is clear that $m_X(n) = 0$, so the correlation function can be written as follows assuming $m > n$.

$$\begin{aligned} R_X(n,m) &= \mathbb{E}[X_n X_m] = \mathbb{E}\left[X_n \cdot (X_n + X_m - X_n)\right] && (15.11)\\ &= \mathbb{E}[X_n^2] + \mathbb{E}\left[X_n \cdot (X_m - X_n)\right] && (15.12)\\ &= n\sigma^2 + \mathbb{E}\left[X_n\right] \cdot \mathbb{E}\left[X_m - X_n\right] \qquad (X_n \perp\!\!\!\perp X_m - X_n) && (15.13)\\ &= n\sigma^2 && (15.14)\\ R_X(n,m) &= \sigma^2 \cdot \min\{m,n\} && (15.15) \end{aligned}$$

■

Here are some properties of the correlation function.

Property 15.6

$$R_X(t_1,t_2) = R_X(t_2,t_1)$$

Property 15.7

$$R_X(t,t) \geq 0 \qquad \forall t \in \mathsf{T}$$

Property 15.8

$$|R_X(t_1,t_2)| \leq \sqrt{\mathbb{E}[X_{t_1}^2] \cdot \mathbb{E}[X_{t_2}^2]}$$

This is a direct corollary of the Cauchy-Schwarz inequality, which you may recall from the linear algebra connection of the expectation on the space of random variables.

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]} \tag{15.16}$$

Proof. For any $\alpha \in \mathbb{R}$, we can write the following.

$$0 \leq \mathbb{E}[(X - \alpha Y)^2] = \mathbb{E}[X^2] - 2\alpha\mathbb{E}[XY] + \alpha^2\mathbb{E}[Y^2] \tag{15.17}$$

If one lets $\alpha = \mathbb{E}[XY]/\mathbb{E}[Y^2]$ and scale by $\mathbb{E}[Y^2]$, we then have the following.

$$0 \leq \mathbb{E}[X^2] \cdot \mathbb{E}[Y^2] - \mathbb{E}[XY]^2 \tag{15.18}$$

Isolating for $\mathbb{E}[XY]$, we obtain the desired inequality. ■

A particularly nice class of correlation functions that are assumed in practice are those with order, or regularity, on the class of functions.

Definition 15.9 (Second Order Process). A random process $\{X_t\}_{t\in\mathsf{T}}$ for which $R_X(t,t) = \mathbb{E}[X_t^2] < \infty$ for all $t \in \mathsf{T}$ is called a **second-order** process.

Aside: An example of a process that is **not** second-order is white noise, which we will discuss later.

For a second-order process, one can show the following properties.

1. $|R_X(t_1,t_2)| < \infty, \forall t_1, t_2 \in \mathsf{T}$
2. $|m_X(t)| \leq \sqrt{\mathbb{E}[X_t^2]} < \infty$

15.3 Stationarity

The concept of a stochastic process as a collection of random variables is too general. To make it useful for applications and to say something meaningful about its properties, we need to make some assumptions on $\{X_t\}_{t\in\mathsf{T}}$. One such assumption that leads to an important subclass of processes is **stationarity**. The intuition behind stationarity is that the "amount" of randomness (formally, the statistics of the process) do not change as function of time, even though the process itself is random.

Example 15.10 Consider the fluctuations of the average daily temperature during summer, in which we observe that it is warm daily. Even if we do not know the exact temperatures, we do not need to check the forecast to know we do not need a jacket. ▪

Aside: One can liken the idea stationarity to time invariance of the system.

We now want to translate this intuition into a rigorous definition.

Definition 15.11 (Strict Stationarity). A random process $\{X_t\}_{t\in\mathsf{T}}$ is said to be **$\mathbf{n^{th}}$ order strict stationary** if $\forall t_1, \cdots, t_n \in \mathsf{T}, \forall \Delta t \in \mathbb{R}$ and $\forall A \subset \mathbb{R}^n$

$$\mathbb{P}\left((X_{t_1+\Delta t}, \cdots, X_{t_n+\Delta t}) \in A\right) = \mathbb{P}\left((X_{t_1}, \cdots, X_{t_n}) \in A\right)$$

Equivalently, this means that the joint distribution of $X_{t_1+\Delta t}, \ldots, X_{t_n+\Delta t}$ does not depend on Δt.

If $\{X_t\}_{t\in\mathsf{T}}$ is n^{th} order strict stationary for all $n \geq 1$, then $\{X_t\}$ is said to be a **strictly stationary stochastic** (SSS) process.

Example 15.12 We show stationarity (or lack thereof) in the aforementioned processes.

Aside: Note the difference between stationarity of the process and stationarity of increments.

1. **Poisson Process:** For $\{X_t\}_{t\geq 0}$ such that $X_t \sim \text{POISSON}(\lambda \cdot t)$, it is not even 1^{st} order strict stationary.

$$X_{t+\Delta t} \sim \text{POISSON}(\lambda \cdot (t + \Delta t)) \tag{15.19}$$

2. **Brownian Motion:** For $\{X_t\}_{t\geq 0}$ such that $X_t \sim \text{NORMAL}(\mu t, \sigma^2 t)$, it is not even 1^{st} order strict stationary.

$$X_{t+\Delta t} \sim \text{NORMAL}(\mu \cdot (t + \Delta t), \sigma^2 \cdot (t + \Delta t)) \tag{15.20}$$

3. **Random Phase:** For $\{X_t\}_{t\in\mathbb{R}}$ such that $X_t = \cos(2\pi f t + \Theta)$, noting that $\Theta \sim \text{UNIFORM}[-\pi, \pi]$ and frequency f, it is a SSS process as $\tilde{\Theta} \sim \text{UNIFORM}[-\pi + 2\pi f \Delta t, \pi + 2\pi f \Delta t]$ is uniformly distributed, and adding a deterministic portion will not affect the distribution.

$$X_{t+\Delta t} = \cos(2\pi f t + 2\pi f \Delta t + \Theta) \tag{15.21}$$
$$= \cos(2\pi f t + \tilde{\Theta}) \qquad (\tilde{\Theta} = 2\pi f \Delta t + \Theta) \tag{15.22}$$

4. **Independent Sum**: For $\{X_n\}_{n\in\mathbb{N}}$ such that X_n is a sum of independent normal random variables.

$$X_n = \sum_{i=1}^{n} Z_i \qquad \{Z_i\}_{i=1}^{n} \sim \text{NORMAL}(0, \sigma^2) \text{ i.i.d} \tag{15.23}$$

We compute $X_{t+\Delta t}$ to show that it is not even 1^{st} order strict stationary.

$$X_{t+\Delta t} \sim \text{NORMAL}(0, (t + \Delta t)\sigma^2) \tag{15.24}$$

▪

Example 15.13 We consider some other trivial examples of SSS processes below.

1. **Same Variable:** For $\{X_t\}_{t\geq 0}$ such that $X_t = X$,
$$\mathbb{P}\left((X_{t_1+\Delta t}, \cdots, X_{t_n+\Delta t}) \in A\right) = \mathbb{P}\left((X, \cdots, X) \in A\right) \tag{15.25}$$
2. **I.I.D:** For $\{X_n\}_{n\in\mathbb{N}}$ i.i.d and sample points $\{n_1, \cdots, n_m\}$,
$$F_{X_{n_1+\Delta t}, \cdots, X_{n_m+\Delta t}}(x_1, \cdots, x_m) = \prod_{i=1}^{n} F_{X_{n_i+\Delta t}}(x_i) = \prod_{i=1}^{n} F(x_i) \tag{15.26}$$

▪

The above examples illustrate that SSS processes are quite simplistic, which is to be expected due to condition being a very strong property with many implications that limit the expressivity of the process.

Theorem 15.14 (Classification of SSS Process). Let $\{X_t\}_{t\in\mathsf{T}}$ be a SSS process.

1. For all $t \in \mathsf{T}$, the random variables are i.i.d. In particular, the mean function does not depend on t.
$$m_X(t) = m$$
2. For all $t_1, t_2 \in \mathsf{T}$, the distribution only depends on the lag $\tau := t_2 - t_1$.
$$\mathbb{P}\left((X_{t_1+\Delta t}, X_{t_2+\Delta t}) \in A\right) \overset{\Delta t = -t_2}{=} \mathbb{P}\left((X_{t_1-t_2}, X_0) \in A\right)$$
In particular, the correlation function depends only on the lag τ.
$$R_X(t_1, t_2) = \mathbb{E}[X_{t_1} X_{t_2}] = f(\tau)$$

From our previous discussion, it is desirable to assume that the process $\{X_t\}_{t\in\mathsf{T}}$ that models some phenomenon of interest is stationary: its analysis is simpler since the probabilistic properties of $\{X_t\}_{t\in\mathsf{T}}$ does not change over time. For example, if we know the past of a stationary process, we can accurately predict its future.

On the other hand, many real-life phenomena are not stationary in the strict sense, as this property is too strong. But even if it is SSS, it is often difficult to verify as we need to ensure that **all** statistics of the process are time invariant. Is all lost? Luckily, no! We can weaken some of the assumptions of an SSS process to get a still useful set of criteria.

Definition 15.15 (Weak Stationarity). A stochastic process $\{X_t\}_{t\in\mathsf{T}}$ is **weak-sense stationary** (WSS) if

1. $m_X(t) = m$ (does not depend on t)
2. $R_X(t_1, t_2) = R_X(\tau)$ (only a function of lag $\tau = t_2 - t_1$)

Aside: We typically denote $R_X(\tau) = \mathbb{E}[X_t X_{t+\tau}]$ as a univariate function for a WSS process.

It is clear that an SSS process is WSS, and even a 2^{nd} order strictly stationary process is WSS. However, the converse is clearly not true, as illustrated in the example below.

Aside: This is analogous to independence implying uncorrelated, but not the converse.

Example 15.16 Let $Z_1, \cdots, Z_n \sim \text{NORMAL}(0, 1)$. We now define $\{X_i\}_{i\in[n]}$ as
$$X_n = \begin{cases} Z_n & \text{if } n \text{ is even} \\ \frac{1}{\sqrt{2}}\left(Z_{n-1}^2 - 1\right) & \text{if } n \text{ is odd} \end{cases} \tag{15.27}$$
Then $\{X_i\}_{i\in[n]}$ is WSS, but not SSS. ▪

Warning 15.17 You should convince yourself that WSS is **not** equivalent to 2^{nd} order SSS, as they look very similar! ▪

Example 15.18 We again investigate whether the aforementioned processes are WSS.

1. **Poisson Process:** For $\{X_t\}_{t\geq 0}$ such that $X_t \sim \text{POISSON}(\lambda \cdot t)$, it is not WSS since
$$m_X(t) = \lambda t \tag{15.28}$$
2. **Random Phase:** For $\{X_t\}_{t\in\mathbb{R}}$ such that $X_t = \cos(2\pi f t + \Theta)$, noting that $\Theta \sim \text{UNIFORM}[-\pi, \pi]$ and frequency f, it is a WSS process automatically due to it being SSS.
$$m_X(t) = 0 \tag{15.29}$$
$$R_X(\tau) = \frac{\cos(2\pi f \tau)}{2} \tag{15.30}$$
3. **Independent Sum**: Consider $\{X_n\}_{n\in\mathbb{N}}$ such that X_n is a sum of independent normal random variables as follows.
$$X_n = \sum_{i=1}^{n} Z_i \qquad \{Z_i\}_{i=1}^{n} \sim \text{NORMAL}(0, \sigma^2) \text{ i.i.d} \tag{15.31}$$
This is not an WSS process, even though $m_X(t) = 0$.
$$R_X(t_1, t_2) = \sigma^2 \cdot \min\{t_1, t_2\} \tag{15.32}$$

▪

We will now discuss various properties of $R_X(\tau)$, if we consider $\{X_t\}_{t\in\mathsf{T}}$ to be a WSS process.

Property 15.19 $R_X(\tau)$ is even, as a result of symmetry of correlation function

$$R_X(\tau) = R_X(-\tau)$$

Aside: $R_X(\tau)$ is sometimes known as an **auto-correlation** function since it is measuring the correlation in lag as opposed to two distinct processes.

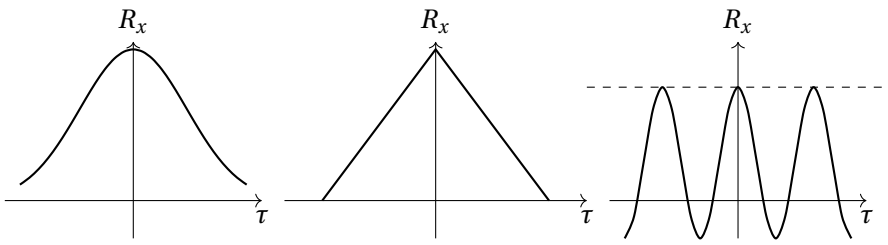
Property 15.20 For a WSS process, the expected power $R_X(0)$ does not depend on t.

$$R_X(0) = \mathbb{E}[X_t^2] \geq 0$$

Property 15.21 For all $\tau \in \mathbb{R}$, $|R_X(\tau)| \leq R_X(0)$ as a result of Cauchy-Schwarz

$$|R_X(\tau)| = |\mathbb{E}[X_t X_{t+\tau}]| \leq \sqrt{\mathbb{E}[X_t^2] \cdot \mathbb{E}[X_{t+\tau}^2]} = R_X(0)$$

Some possible shapes of $R_X(\tau)$ are pictured as follows.



16. Applications of General Stochastic Processes

Last time we introduced an important concept of **stationary** stochastic processes, namely WSS and SSS processes.

- A stochastic process $\{X_t\}_{t\in\mathsf{T}}$ is **SSS** if $\forall n \geq 1, \forall t_1, \cdots, t_n \in \mathsf{T}, \forall \Delta t \in \mathbb{R}$, and $\forall A \subset \mathbb{R}^n$,
$$\mathbb{P}\left((X_{t_1+\Delta t}, \cdots, X_{t_n+\Delta t}) \in A\right) = \mathbb{P}\left((X_{t_1}, \cdots, X_{t_n}) \in A\right)$$
- A stochastic process $\{X_t\}_{t\in\mathsf{T}}$ is **WSS** if we have that **1**) $m_X(t) = \mathbb{E}[X_t] = c$, and **2**) $R_X(t, t+\tau) = R_X(\tau)$.

We showed that for a general process,
$$\text{SSS} \implies \text{WSS} \quad \text{but} \quad \text{WSS} \not\implies \text{SSS}$$
There is an important exception to this rule.

Agenda:
1. Gaussian Processes
2. Interpreting and Estimating $R_X(\tau)$
3. Power Spectral Density

16.1 Gaussian Processes

Gaussianity was an important part of studying random variables before, and this manifests in the study of stochastic processes as well.

Definition 16.1 (Gaussian Process). A stochastic process $\{X_t\}_{t\in\mathsf{T}}$ is said to be a **Gaussian stochastic process** if, for any sequence of $t_1, \cdots, t_n \in \mathsf{T}$, $(X_{t_1}, \cdots, X_{t_n})^\top$ is a Gaussian random vector. Equivalently, $X_{t_1}, \cdots, X_{t_n}$ are jointly normal or any linear combination is normal.
$$\sum_{i=1}^n \alpha_i X_{t_i} \sim \text{NORMAL}$$

We consider an example of such a process and some nice properties.

Example 16.2 Let $\{X_t\}_{t\in\mathbb{R}}$ be a zero-mean, Gaussian WSS process with $R_X(\tau) = \exp(-\tau^2)$. We can now derive some quantities of interest as follows.

To determine $\mathbb{P}(X_t < a)$, we note $\mathbb{E}[X_t] = 0$ and $\mathbb{V}[X_t] = R_X(0) = \exp(0) = 1$. Thus, $X_t \sim \text{NORMAL}(0, 1)$, which means that $\mathbb{P}(X_t < a) = \Phi(a)$.

To determine $\mathbb{P}(X_t + X_s < a)$, we note that $X_t + X_s \sim \text{NORMAL}$ and $\mathbb{E}[X_s + X_t] = 0$. For the variance, we have the following simplification.
$$\mathbb{V}[X_s + X_t] = \mathbb{V}[X_t] + \mathbb{V}[X_s] + \text{Cov}[X_t, X_s] = 2 + \exp(-(t-s)^2) \tag{16.1}$$
Thus, $Y := X_s + X_t \sim \text{NORMAL}(0, 2 + \exp(-(t-s)^2))$, which means $\mathbb{P}(Y < a)$ is as follows.
$$\mathbb{P}(Y < a) = \Phi\left(\frac{a}{\sqrt{2 + \exp(-(t-s)^2)}}\right) \tag{16.2}$$
∎

An important property of Gaussian processes is as follows.

Theorem 16.3 (Stationarity of Gaussian Process). Let $\{X_t\}_{t\in\mathsf{T}}$ be a Gaussian process. If $\{X_t\}_{t\in\mathsf{T}}$ is WSS, it is also SSS.

Aside: This theorem is analogous to uncorrelated, jointly normal random variables being independent.

Proof. We ultimately seek to show that for all $t_1, \cdots, t_n \in \mathsf{T}$ and any Δt, the joint CDF of $\mathbf{u} = (X_{t_1}, \cdots, X_{t_n})^\top$ and the joint CDF of $\mathbf{v} = (X_{t_1+\Delta t}, \cdots, X_{t_n+\Delta t})^\top$ are the same. Since $\mathbf{u}$ and $\mathbf{v}$ are Gaussian vectors, it is enough to show that their mean vectors and covariance matrices are identical.

$$\mathbb{E}[\mathbf{u}] = \mathbb{E}[\mathbf{v}] = (m, \cdots, m)^\top \in \mathbb{R}^n \qquad m = m_X(t) = \mathbb{E}[X_t] \tag{16.3}$$

$$(\Sigma_{\mathbf{v}})_{ij} = \mathrm{Cov}(X_{t_i+\Delta t}, X_{t_j+\Delta t}) = R_X(t_i - t_j) - m^2 = (\Sigma_{\mathbf{u}})_{ij} \tag{16.4}$$

■

As one can see, Gaussianity is an important principle in all of probability theory that allows us to make some powerful assumptions. An application of Gaussian processes is in **supervised learning**.

To recall the framework of supervised learning, we seek to pin down an underlying function $f(\mathbf{x}; \theta)$ that sends a feature vector $\mathbf{x}$ to a prediction y, which is trained using some dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ where y_i is assumed to be computed as follows. We seek to compute the predictive **distributions** on test points $\mathcal{T} = \{\tilde{\mathbf{x}}_i\}_{i=1}^M$.

$$y_i = f(\mathbf{x}_i; \theta) + \epsilon_i \qquad \epsilon_i \sim \textsc{normal}(0, \sigma_\epsilon^2) \tag{16.5}$$

In the typical Bayesian approach, one attempts to construct a posterior distribution over the parameter space θ and find the most probable value, known as the MAP point. If we do not want to fix θ, then how can one solve this problem?

A powerful dual perspective is to place a distribution on the **functions**. One can parameterize $f(\mathbf{x})$ as a Gaussian process. The kernel K is the driving force of this assumption, allowing us to model various classes of functions depending on the behavior of our data.

$$f(\mathbf{x}) \sim \mathrm{GP}(\mu(\mathbf{x}), K(\mathbf{x}, \mathbf{x}')) \tag{16.6}$$

Thus, our prior distribution is simply a trajectory of this Gaussian process, and conditioning on the data we see yields a posterior Gaussian process, from which we can sample from to obtain our posterior function. The details of this are far beyond the scope of this class, but our principles of conditioning should help you make sense of this!

Aside: For more details, Professor Owhadi teaches a wonderful class (ACM 118) on kernel methods in general and applications.

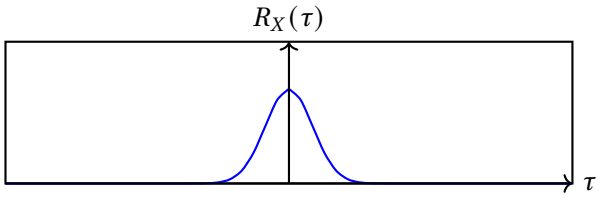
Since functions live in infinite-dimensional spaces, assumptions are often made by parameterizing them (i.e. linear, polynomial, etc). However, Gaussian process regression is non-parametric, in that the choice of kernel allows us to model these infinitely many functions without limiting the model.

16.2 Interpreting and Estimating $R_X(\tau)$

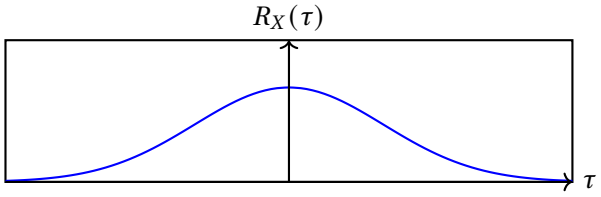
Let $\{X_t\}_{t\in\mathsf{T}}$ be a WSS process. In Lecture 16, we showed that the correlation function only depends on the time lag τ, or that $R_X(\tau) = \mathbb{E}[X_t X_{t+\tau}]$. Recall the properties of this function as follows:

1. $R_X(-\tau) = R_X(\tau)$
2. $R_X(0) \geq 0$
3. $|R_X(\tau)| \leq R_X(0)$

Intuitively, if $m_X(t) = 0$ and $R_X(\tau)$ decays quickly with τ, then it is reasonable to infer that X_t and $X_{t+\tau}$ become quickly uncorrelated as τ increases. More concretely, $R_X(\tau)$ controls the probability of a change in the stochastic process.



$X_t, X_{t+\tau}$ uncorrelated even for small τ



$X_t, X_{t+\tau}$ correlated even for large τ

We can quantify this "decay" through the following theorem.

Theorem 16.4 (Correlation Decay of WSS). If $\{X_t\}_{t\in\mathsf{T}}$ is WSS, then

$$\mathbb{P}(|X_{t+\tau} - X_t| \geq a) \leq \frac{2 \cdot (R_X(0) - R_X(\tau))}{a^2}$$

Aside: This is known as a **tail** probability, and is common in the study of large deviations.

Proof. Using Chebyshev's inequality, this tail probability simplifies as follows.

$$\mathbb{P}(|X_{t+\tau} - X_t| \geq a) \leq \frac{\mathbb{V}[X_{t+\tau} - X_t]}{a^2} = \frac{\mathbb{E}[(X_{t+\tau} - X_t)^2] - \mathbb{E}[X_{t+\tau} - X_t]^2}{a^2} \tag{16.7}$$

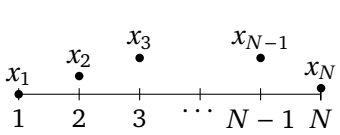
$$= \frac{\mathbb{E}[X_{t+\tau}^2] - 2\mathbb{E}[X_t X_{t+\tau}] + \mathbb{E}[X_t^2]}{a^2} \tag{16.8}$$

$$= \frac{R_X(0) - 2R_X(\tau) + R_X(0)}{a^2} = \frac{2 \cdot (R_X(0) - R_X(\tau))}{a^2} \tag{16.9}$$

■

From the above interpretation of $R_X(\tau)$, it is clear that it is a crucial characteristic of a WSS process. But in practical problems, it is not given and usually requires estimation from data.

We consider the two classes of data to be modeled from a WSS process, either discrete points or a single continuous trajectory of a realization of the process.





For the **former** case, we can estimate $R_X(n)$ through the random variable below.

$$Y_n = \frac{1}{N-n} \cdot \sum_{i=1}^{N-n} X_i X_{i+n} \tag{16.10}$$

Taking the expectation, we have the following.

$$\mathbb{E}[Y_n] = \frac{1}{N-n} \cdot \sum_{i=1}^{N-n} \mathbb{E}[X_i X_{i+n}] = \frac{1}{N-n} \cdot \sum_{i=1}^{N-n} R_X(n) = R_X(n) \tag{16.11}$$

This then motivates us to empirically estimate $R_X(n)$ via $\{x_i\}_{i=1}^N$ through the following.

$$R_X(n) = \frac{1}{N-n} \cdot \sum_{i=1}^{N-n} x_i \cdot x_{i+n} \tag{16.12}$$

If we instead observe the data continuously, as in $\{x_t\}_{t\in[0,T]}$, then an analogous random variable can be constructed.

$$Y_\tau = \frac{1}{T-\tau} \cdot \int_0^{T-\tau} X_{t+\tau} X_t \, \mathrm{d}t \tag{16.13}$$

Stochastic integration is far outside the scope of this class (see SS 214), but we can circumvent the details by considering the Riemann sum approximation of this integral.

1. Fix $m > 0$ and let $\Delta t = \frac{T-\tau}{m}$, such that $t_k = k \cdot \Delta t$ for $k = 0, \cdots, m-1$
2. Construct the Riemann sum approximation of Y_τ

$$Y_{\tau,m} = \frac{1}{T-\tau} \cdot \sum_{k=0}^{m-1} X_{t_k} \cdot X_{t_k+\tau} \cdot \Delta t \tag{16.14}$$

We remark that the following limit holds true.

$$\lim_{m\to\infty} Y_{\tau,m} = Y_\tau \tag{16.15}$$

Aside: Barring nuances, this heuristic discretization is surprisingly reliable when numerically solving SDEs.

Since integration is a linear operation, we can once again compute the expectation as follows for $m \gg 1$.

$$\mathbb{E}[Y_\tau] \approx \mathbb{E}[Y_{\tau,m}] = \frac{1}{T-\tau} \cdot \sum_{k=0}^{m-1} \mathbb{E}[X_{t_k} \cdot X_{t_k+\tau}] \cdot \Delta t \tag{16.16}$$

$$= \frac{1}{T-\tau} \cdot \sum_{k=0}^{m-1} R_X(\tau) \cdot \Delta t \approx \frac{1}{T-\tau} \cdot \int_0^{T-\tau} R_X(\tau) \, \mathrm{d}t = R_X(\tau) \tag{16.17}$$

$$\tag{16.18}$$

This in turn suggests that we can estimate $R_X(\tau)$ through the following formula.

$$R_X(\tau) = \frac{1}{T-\tau} \cdot \int_0^{T-\tau} x_t \cdot x_{t+\tau} \, \mathrm{d}t \tag{16.19}$$

16.3 Power Spectral Density

Let $\{X_t\}_{t\in\mathsf{T}}$ be a WSS process with the correlation function $R_X(\tau)$.

Definition 16.5 (Power Spectral Density). The **power spectral density** (PSD) of $\{X_t\}_{t\in\mathsf{T}}$ is the Fourier transform $\mathcal{F}[\cdot]$ of $R_X(\tau)$.

$$S_X(k) = \mathcal{F}[R_X(\tau)] = \begin{cases} \int_{-\infty}^{\infty} R_X(\tau)\cdot\exp(-2\pi i\tau k)\,\mathrm{d}\tau & \text{time-continuous} \\ \sum_{\tau=-\infty}^{\infty} R_X(\tau)\cdot\exp(-2\pi i\tau k), |k| < \frac{1}{2} & \text{time-discrete} \end{cases}$$

The correlation function $R_X(\tau)$ can be recovered from the PSD $S_X(k)$ by taking the inverse Fourier transform.

$$R_X(\tau) = \mathcal{F}^{-1}[S_X(k)] = \begin{cases} \int_{-\infty}^{\infty} S_X(k)\cdot\exp(2\pi i\tau k)\,\mathrm{d}k & \text{time-continuous} \\ \int_{-1/2}^{1/2} S_X(k)\cdot\exp(2\pi i\tau k)\,\mathrm{d}k & \text{time-discrete} \end{cases} \tag{16.20}$$

Aside: One can extend the PSD for non-WSS processes, albeit with much more difficulty.

We can show that the PSD makes physical sense as $S_X(k)$ can be interpreted as a power density, where we recall that $\mathbb{E}[X_t^2]$ is the expected power at time t. As seen below, we are averaging the power across all possible frequencies.

$$\mathbb{E}[X_t^2] = R_X(0) = \int_{-\infty}^{\infty} S_X(k)\cdot\exp(-2\pi i\tau k)\,\mathrm{d}k\Big|_{\tau=0} = \int_{-\infty}^{\infty} S_X(k)\,\mathrm{d}k \tag{16.21}$$

We can easily show the following properties of the PSD through the definition of the Fourier transform as follows.

1. $\forall k, S_X(k) \in \mathbb{R}$ (real-valued)
2. $S_X(-k) = S_X(k)$ (even)
3. $S_X(f) \geq 0$ (non-negative)

The first two properties can be show by expanding the exponential in terms of trigonometric functions as follows. In the final step, we use the fact that $R_X(\tau)$ is even to conclude that $S_X(k)$ is also even.

$$S_X(k) = \int_{-\infty}^{\infty} R_X(\tau)\cdot\cos(-2\pi\tau k)\,\mathrm{d}\tau + i\cdot\int_{-\infty}^{\infty} R_X(\tau)\cdot\underbrace{\sin(-2\pi\tau k)}_{\text{odd}}\,\mathrm{d}\tau \tag{16.22}$$

$$= \int_{-\infty}^{\infty} R_X(\tau)\cdot\underbrace{\cos(2\pi i\tau k)}_{\text{even}}\,\mathrm{d}\tau \in \mathbb{R} \tag{16.23}$$

Aside: Recall that integrating an odd function across symmetric bounds yields 0.

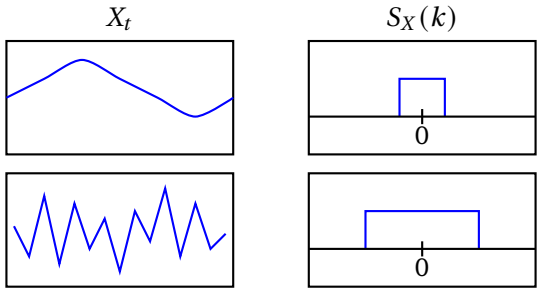
More surprisingly, these conditions completely characterize the PSD. This gives us the following result.

Theorem 16.6 (Existence of PSD). Let $g(k)$ be any real-valued, even, and non-negative function. Then there exists a WSS process $\{X_t\}_{t\in\mathsf{T}}$ such that $S_X(k) = g(k)$.

Aside: The theorem can be strengthened to say that any such function has an associated **Gaussian** WSS process.

In turn, this theorem now allows us to check if a claimed correlation function $\hat{R}(\tau)$ is indeed valid, equivalently if there exists a WSS process $\{X_t\}_{t\in\mathsf{T}}$ such that $\hat{R}(\tau) = R_X(\tau)$. One can do this by computing $\mathcal{F}[\hat{R}_X(\tau)]$ and checking the necessary conditions.

One can use the intuitive definition of the Fourier transform as a signal decomposition of sorts to remark on what the PSD implies about the process $\{X_t\}_{t\in\mathsf{T}}$ itself. The smoother the sample paths of $\{X_t\}_{t\in\mathsf{T}}$ are, the lower the frequency content of $S_X(k)$ is. Conversely, the "wigglier" the sample path is, the higher the frequency content of $S_x(k)$ is. Consider the heuristic visual below for two different types of processes.



A natural limiting process in the frequency domain arises, defined below.

Definition 16.7 (Bandlimited White Noise). A zero-mean WSS process with $S_X(k)$ defined as

$$S_X(k) = \begin{cases} \frac{n}{2} & k \in [-f_0, f_0] \\ 0 & \text{elsewhere} \end{cases}$$

is called a **bandlimited** white noise.

To find the correlation function, we compute the inverse Fourier transform.

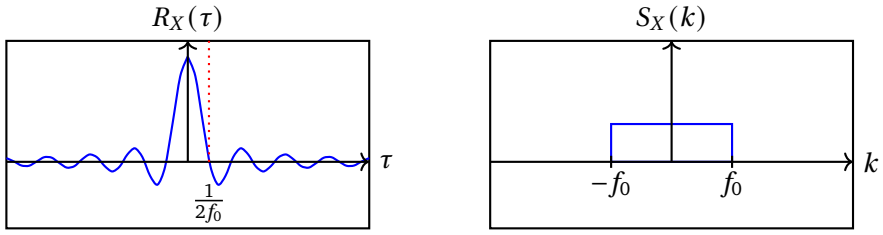
$$R_X(\tau) = \int_{-\infty}^{\infty} S_X(k)\cdot\exp(2\pi i\tau k)\,\mathrm{d}k = \int_{-f_0}^{f_0} \frac{n}{2}\cdot\exp(2\pi i\tau k)\,\mathrm{d}k \tag{16.24}$$

$$= \frac{n}{4\pi i\tau}\cdot\exp(2\pi i\tau k)\Big|_{-f_0}^{f_0} = \frac{n\cdot(\exp(2\pi i\tau f_0)-\exp(-2\pi i\tau f_0))}{4\pi i\tau} \tag{16.25}$$

$$= \frac{n\cdot 2i\sin(2\pi\tau f_0)}{4\pi i\tau} = \frac{nf_0\cdot\sin(2\pi\tau f_0)}{2\pi f_0\tau} = nf_0\cdot\operatorname{sinc}(2\pi f_0\tau) \tag{16.26}$$

Aside: As $f_0\to\infty$, we recover the prototypical white noise.

We consider the plots of these two functions as follows.



The PSD is a fundamental concept in signal processing, and you will learn about its applications in more advanced EE courses. One statistical application of this is:

1. Estimate $R_X(\tau)$ from data as follows.
2. Estimate $S_X(k)$ from $R_X(\tau)$.
3. Generate realizations (sample paths) of the stochastic process with $S_X(k)$ and use to estimate quantities of interest.

Bibliography

[Dia07] P. Diaconis. “Dynamics bias in the coin toss”. In: (2007).

[Kel86] J. Keller. “The Probability of Heads”. In: (1986).