EECS 126 Notes

I took these notes live during the Spring 2021 iteration of EECS 126, which was taught by Prof. Thomas Courtade. While these originally started out as just scribed lecture notes, I went back and added some descriptions, facts and examples on my own, in an effort to make this source slightly more comprehensive for myself.

Contents

1 Lecture 1 **4**
1.1 Axioms . . . 4
1.2 Consequences of the axioms . . . 5
1.3 Questions to Ask . . . 6

2 Lecture 2 **7**
2.1 Conditional Probability . . . 7
2.2 Independence . . . 9
2.3 Conditional Independence . . . 10
2.4 Random Variables Intro . . . 10

3 Lecture 3 **11**
3.1 Random Variables Intro Cont. . . . 11
3.2 Discrete Random Variables . . . 11
3.3 Joint Distributions . . . 12
3.4 Expectation . . . 12

4 Lecture 4 **16**
4.1 Variance . . . 16
4.2 Common Examples of Discrete Distributions . . . 18
4.3 Conditional Distributions . . . 20

5 Lecture 5 **22**
5.1 Conditional Expectation . . . 22
5.2 Continuous Random Variables . . . 23
5.3 Independence . . . 24
5.4 Common Examples of Continuous Random Variables . . . 25
5.5 Expectation and Variance of Continuous Random Variables . . . 25
5.6 Conditioning . . . 26

6 Lecture 6 **28**
6.1 Review . . . 28
6.2 Memoryless Property . . . 28
6.3 Gaussian Random Variable . . . 29

7 Lecture 7 **32**
7.1 Conditional Variance . . . 32
7.2 Derived Distributions . . . 32
7.3 Order Statistics . . . 34
7.4 Convolutions . . . 35

8 Lecture 8 **37**
8.1 Moment Generating Function . . . 37
8.2 Characteristic Function . . . 38
8.3 Concentration Inequalities . . . 39
8.4 Weak Law of Large Numbers . . . 39

9 Lecture 9 **41**
9.1 Chernoff Bounds . . . 41
9.2 Convergence of Random Variables . . . 41
9.3 Strong Law of Large Numbers . . . 43
9.4 Central Limit Theorem . . . 44

10 Lecture 10 **45**
10.1 Information Theory . . . 45
10.2 Source Coding (i.e. Compression) . . . 45
10.3 Properties of Entropy . . . 46
10.4 Asymptotic Equipartition Theorem . . . 48

11 Lecture 11 **49**
11.1 Applying AEP to Source Coding . . . 49
11.2 Information Transmission (Channel Coding) . . . 50

12 Lecture 12 **53**
12.1 Information Transmission (Channel Coding) Cont. . . . 53
12.2 Markov Chains . . . 54

13 Lecture 13 **57**
13.1 Classification of States . . . 57
13.2 Class Properties . . . 57
13.3 Long Term Behavior . . . 59

14 Lecture 14 **61**
14.1 Recap of Lecture 13 . . . 61
14.2 Reversibility . . . 61
14.3 First Step Analysis . . . 62

15 Lecture 15 **64**
15.1 First Step Analysis Cont. . . . 64
15.2 Poisson Processes . . . 65
15.3 Conditional Distribution of Arrivals . . . 66

16 Lecture 16 **68**
16.1 Recap . . . 68
16.2 Merging and Splitting/Thinning . . . 68
16.3 Random Incidence Paradox . . . 69
16.4 Continuous Time Markov Chains . . . 70

17 Lecture 17 **72**
17.1 CTMCs Cont. . . . 72
17.2 Stationary Distributions of CTMCs . . . 73

18 Lecture 18 **76**
18.1 First Step Analysis . . . 76
18.2 Uniformization . . . 76
18.3 Random Graphs . . . 78

19 Lecture 19 **80**
19.1 Random Graphs Cont. . . . 80
19.2 Connectivity Threshold . . . 80
19.3 Statistical Inference . . . 81

20 Lecture 20 **84**
20.1 Statistical Inference Cont. . . . 84
20.2 Binary Hypothesis Testing . . . 84

21 Lecture 21 **87**
21.1 Neyman-Pearson Lemma Proof . . . 87
21.2 Estimation . . . 88

22 Lecture 22 **91**
22.1 Linear Estimation . . . 91
22.2 Connection to Linear Regression . . . 92
22.3 Geometry of Linear Estimation . . . 93
22.4 Orthogonality Principle . . . 94

23 Lecture 23 **96**
23.1 Orthogonality Principle Cont. . . . 96
23.2 LLSE Error . . . 96
23.3 Applications of the Orthogonality Principle (MMSE) . . . 97
23.4 MMSE Error . . . 97
23.5 Online estimation . . . 98

24 Lecture 24 **99**
24.1 Gram Schmidt for Random Variables . . . 99
24.2 Jointly Gaussian Random Variables . . . 100

25 Lecture 25 **103**
25.1 Jointly Gaussians Random Variables Cont. . . . 103
25.2 Kalman Filtering . . . 103
25.3 Proof of correctness of the scalar KF . . . 104

1 Lecture 1

Definition 1.1: Probability Space
A probability space is a triple $(\Omega, \mathcal{F}, P)$ where

- Ω is the set of all "samples"
- $\mathcal{F}$ is the family of subsets (called "events") of Ω
- P is a probability measure

We say that the sample space Ω is the set of all possible sample points $\omega \in \Omega$.

1.1 Axioms

Definition 1.2: σ-Algebra
σ-algebras are sets of countable events where complements/intersections/unions of events are also events.

We will make the assumption that $\mathcal{F}$ is a "σ-algebra" containing Ω itself. Then, the measure $P : \mathcal{F} \to [0, 1]$ assigns probabilities to events in $\mathcal{F}$. Probability measures must obey the Kolmogorov axioms:

Definition 1.3: Kolmogorov Axioms

1. $P(A) \geq 0, \forall A \in \mathcal{F}$
2. $P(\Omega) = 1$
3. If $A_1, A_2, \cdots \in \mathcal{F}$ is a countable sequence and all A_i are disjoint, then $P\left(\bigcup_{i\geq 1} A_i\right) = \sum_{i\geq 1} P(A_i)$

Note 1.1
Why countable? Consider the following:

$$1 = P([0,1]) = \sum_{x\in[0,1]} P(\{x\}) = 0$$

The notation above is applying the probability measure on the interval $[0, 1]$, which is an uncountable set, and $P(x) = 0$ for any continuous distribution.

Example 1.1
Flip a coin with bias p

$$\begin{aligned}\Omega &= \{H, T\}\\ \mathcal{F} &= \{\emptyset, H, T, \{H, T\}\}\\ P(H) &= p\\ P(T) &= 1 - p\\ P(\emptyset) &= 0\\ P(\{H, T\}) &= 1\end{aligned}$$

The choice given above is a valid probability space since it obeys all the axioms.

Example 1.2
Here is a different possibility: define the events

$$A = \{\text{all configurations of atoms such that the coin lands heads}\}$$
$$B = \{\text{all configurations of atoms such that the coin lands tails}\}$$

Then,

$$\begin{aligned} \Omega &= A \cup B \\ \mathcal{F} &= \{\emptyset, A, B, \{A \cup B\}\} \\ P(A) &= p \\ P(B) &= 1 - p \\ P(\emptyset) &= 0 \\ P(A \cup B) &= 1 \end{aligned}$$

The choice above also obeys all the axioms so it is a valid choice for a probability space as well.

Example 1.3
Flip 2 coins, biased to heads with probability p and q respectively, such that

$$\begin{aligned} \Omega &= \{HH, HT, TH, TT\} \\ A &= \{HH, HT\} \\ B &= \{TH, TT\} \\ \mathcal{F} &= \{\emptyset, A, B, \Omega\} \\ P(A) &= p \\ P(B) &= 1 - p \end{aligned}$$

where A and B are the events that the first toss is heads and tails respectively.

Example 1.4
Another valid alternative to the configuration given above is

$$\begin{aligned} \mathcal{F} &= 2^{\Omega} = 2^{\{H,T\}} \\ P(HH) &= pq \\ P(HT) &= p(1-q) \\ P(TH) &= (1-p)q \\ P(TT) &= (1-p)(1-q) \end{aligned}$$

1.2 Consequences of the axioms

These simple, yet powerful, statements are a direct result of the Kolmogorov axioms:

- If Ω is countable, and each $\omega \in \Omega$ is a sample point, then $\sum_{\omega \in \Omega} P(\omega) = 1$
- $P(A^c) = 1 - P(A)$

  Proof: $1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof: $P(A \cup B) = P(A \setminus (A \cap B)) + P(B) = P(A) - P(A \cap B) + P(B)$

- If $\mathcal{F}$ is closed under countable unions and complements, then it is also closed under countable intersections.

  Proof: If $A_1, A_2, \cdots \in \mathcal{F}$ (for countably many A_i), then

$$A_1^c, A_2^c, \ldots \in \mathcal{F} \qquad \text{closed under complements}$$

$$\bigcup_{i \geq 1} A_i^c \in \mathcal{F} \qquad \text{closed under countable unions}$$

$$\left(\bigcup_{i \geq 1} A_i^c \right)^c = \bigcap_{i \geq 1} A_i \in \mathcal{F} \qquad \text{closed under complements}$$

1.3 Questions to Ask

Probability is an extremely powerful tool that can be used in a wide array of disciplines. For example, a mathematician might ask: given a model $(\Omega, \mathcal{F}, P)$, what can I say about the outcomes of different experiments? On the other hand, a statistician might ask: given the outcomes of an experiment, what is a good model $(\Omega, \mathcal{F}, P)$ that could have produced them? Whereas engineers and scientists are generally more concerned about choosing the right model $(\Omega, \mathcal{F}, P)$ to capture or approximate some problem they are facing and analyzing it to draw valuable insights for system design, control, experiments, etc.

2 Lecture 2

2.1 Conditional Probability

In applying probability, we often want to compute the probability of pretty complicated events. One tool to help us is the law of total probability.

Theorem 2.1: Law of Total Probability
If $A_1, A_2 \dots$ partition Ω and the A_i's are disjoint, then for any event B,

$$P(B) = \sum_{i \geq 1} P(B \cap A_i)$$

Proof: By simply applying Kolmogorov's axioms,

$$\begin{aligned} P(B) &= P(B \cap \Omega) \\ &= P\left(B \cap \bigcup_{i \geq 1} A_i\right) \\ &= P\left(\bigcup_{i \geq 1} (A_i \cap B)\right) \\ &= \sum_{i \geq 1} P(A_i \cap B) \end{aligned}$$

Example 2.1
Suppose you are a part of surveillance testing for COVID. A person is infected with probability 0.02 and not infected with probability $1 - 0.02 = 0.98$. An infected person (who will always test positive) is symptomatic with probability 0.7 and asymptomatic with probability 0.3. A non-infected person tests positive (false positive test) with probability 0.02 and negative (true negative test) with probability 0.98. What is the probability of you testing positive yet being asymptomatic? Using the law of total probability,

$$\begin{aligned} P(\{+, \text{asymptomatic}\}) &= P(\{+, \text{asymptomatic}\} \cap \{\text{false positive result}\}) \\ &\quad + P(\{+, \text{asymptomatic}\} \cap \{\text{true positive result}\}) \\ &\quad + P(\{+, \text{asymptomatic}\} \cap \{\text{negative result}\}) \\ &= 0.98 \times 0.02 + 0.02 \times 0.3 + 0 \\ &= 0.0256 \end{aligned}$$

Definition 2.1: Conditional Probability
If B is an event with $P(B) > 0$, then the conditional probability of A given B is

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

Conceptually, $A \mid B$ is the event A after taking into account the knowledge that event B has occurred.

Theorem 2.2: Bayes Rule
If events A and B have positive probability, then

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

Proof: This follows directly from the definition of conditional probability:

$$\begin{aligned}\frac{P(B \mid A)P(A)}{P(B)} &= \frac{P(A \cap B)}{P(A)} \cdot \frac{P(A)}{P(B)} \\ &= \frac{P(A \cap B)}{P(B)} \\ &= P(A \mid B)\end{aligned}$$

Example 2.2

We roll 2 standard die, and the sum is 10. What is the probability that the first roll was a 4? Let A be the event that the first roll is a 4 and B be the event the sum of rolls is 10. Then,

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)} = \frac{P(\{4,6\})}{P(\{4,6\}) + P(\{5,5\}) + P(\{6,4\})} = \frac{1}{3}$$

Example 2.3

If $A_1, \ldots, A_n$ are disjoint events that partition Ω, then the law of total probability can be rewritten as

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$$

Substituting this into Bayes Rule,

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}$$

This is called the extended Bayes Rule.

Theorem 2.3: Chain Rule

For conditional probability, we can generally decompose

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) \times P(A_2 \mid A_1) \times \cdots \times P(A_n \mid A_1, \ldots, A_{n-1})$$

This is called the product rule or the chain rule of probability and it can be proven using induction.

Example 2.4: Birthday Paradox

Let n people be in a room. What is the probability that no two people share a common birthday? For $i = 1, \ldots, n$, let

$$A_i = \{\text{person i does hare a birthday with any person } j = 1, \ldots, i-1\}$$

Then,

$$\begin{aligned}\mathbb{P}[\text{no common birthday}] &= \mathbb{P}\left[\bigcap_{i \geq 1}^{n} A_i\right] \\ &= \prod_{i=1}^{n} \mathbb{P}[A_i \mid A_1 \cap \cdots \cap A_{i-1}]\end{aligned}$$

Observe that $\mathbb{P}[A_1 \mid A_1 \cap \cdots \cap A_{n-1}] = \frac{365-(i-1)}{365}$ since $i-1$ choices have been taken. Thus,

$$\begin{aligned}\mathbb{P}\left[\bigcap_{i \geq 1}^{n} A_i\right] &= \prod_{i=1}^{n} \left(1 - \frac{i-1}{365}\right) \\ &\leq \prod_{i=1}^{n} e^{-\frac{i-1}{365}}\end{aligned}$$

$$= e^{-\frac{\sum_{i=1}^{n}(i-1)}{365}}$$
$$= e^{-\frac{\binom{n}{2}}{365}}$$

Thus, $\mathbb{P}[\text{2 people share a common birthday}] \geq 1 - e^{-\frac{\binom{n}{2}}{365}}$. When $n = 23$, this quantity is ≈ 0.5!

Note 2.1
Generally, conditioning on some event will restrict the overall sample space to just that event. Intuitively, this is because we are only concerned with the sample points that are a part of that event. One consequence of this is the modified law of total probability for conditional probabilities:

$$P(B \mid C) = \sum_{i \geq 1} P(B \mid A_i \cap C) P(A_i \mid C)$$

where all A_i are disjoint subsets of Ω but Ω itself is conditioned on C. This notion also holds for other equations like Union Bound, Bayes Rule, etc.

2.2 Independence

Definition 2.2: Independence
Events A and B are independent if and only if $P(A \cap B) = P(A)P(B)$

Note 2.2
Sometimes this technical definition of independence does not agree with intuition but it usually does. However, it is still the most useful and commonly applicable definition of independence.

Note 2.3
If $P(B) > 0$, then A, B being independent is equivalent to saying $P(A \mid B) = P(A)$. Basically, knowing B occurred tells us nothing about A's occurrence.

Note 2.4
In general, events $A_1, A_2, \ldots$ are independent iff

$$P\left(\bigcap_{i \in S} a_i\right) = \prod_{i \in S} P(A_i)$$

for all finite subsets of indices S.

Note 2.5
If $A_1, A_2, \ldots$ are independent, then so are $B_1, B_2, \ldots$ where each B_i is either A_i or A_i^c.

Example 2.5
Consider infinite sequence of independent fair coin tosses. What does this mean, precisely?

$$A_i = \{\text{toss } i \text{ is heads}\}$$
$$A_i^c = \{\text{toss } i \text{ is tails}\}$$

$$P\left(\bigcap_{i=1}^{n} \text{toss } i \text{ has outcome } x_i\right) = \prod_{i=1}^{n} P(\text{toss } i \text{ has outcome } x_i)$$
$$= \prod_{i=1}^{n} \frac{1}{2}$$
$$= 2^{-n}$$

Note 2.6
Pairwise independence (both events in all possible pairs of events are independent) does not imply mutual independence (all events are independent of each other) and vice versa. The standard counterexample goes as follows: flip two fair coins and let A be the event that the first coin lands heads, B be the event that the second coin lands heads and C be the event that both tosses are different.

Note 2.7
Mutually exclusive events are not independent: one event occurring prevents the other from occurring.

2.3 Conditional Independence

Definition 2.3: Conditional Independence
If events A, B and C satisfy $P(A \cap B \mid C) = P(A \mid C)P(B \mid C)$ for $P(C) > 0$, then we say that A and B are conditionally independent on C.

Example 2.6
There are 2 coins with bias p and q respectively. Pick a coin at random and flip twice. Let H_i be the event that toss i lands heads. Then,

$$P(H_i) = \frac{1}{2}p + \frac{1}{2}q$$
$$P(H_1 \cap H_2) = \frac{1}{2}p^2 + \frac{1}{2}q^2$$

Although H_1 and H_2 are not independent, they are conditionally independent on $C = \{\text{coin } p \text{ is chosen}\}$.

2.4 Random Variables Intro

A random variable is a function $X : \Omega \to \mathbb{R}$ with the property that

$$\{\omega \in \Omega : X(\omega) \leq \alpha\} \in \mathcal{F}, \forall \alpha \in \mathbb{R}$$

This means that we can compute the probability $P(X \leq \alpha) = P(\{\omega \in \Omega : X(\omega) \leq \alpha\})$ for all $\alpha \in \mathbb{R}$. The set above is a valid event.

3 Lecture 3

3.1 Random Variables Intro Cont.

Definition 3.1: Random Variables
For a probability space $(\Omega, \mathcal{F}, P)$, a random variable is a function $X : \Omega \to \mathbb{R}$ with the property that

$$\{\omega : X(\omega) \leq \alpha\} \in \mathcal{F}, \forall \alpha \in \mathbb{R}$$

This ensures that we can compute

$$\mathbb{P}[X \leq \alpha] = P(\{\omega : X(\omega) \leq \alpha\})$$

We can also compute $\mathbb{P}[X \in B]$ for pretty much at set B you will encounter. Why?

- $\mathcal{F}$ is closed under complements, $\{\omega : X(\omega) > \alpha\} \in \mathcal{F}$ for all $\alpha \in \mathbb{R}$.
- $\mathcal{F}$ is closed under intersections, $\{\omega : \alpha < X(\omega) \leq \beta\} \in \mathcal{F}$ for all $\alpha < \beta \in \mathbb{R}$.
- $\mathcal{F}$ is closed under countable unions, $\{\omega : \alpha < X(\omega) < \beta\} \in \mathcal{F}$ i.e., $\bigcup_{n \geq 1} \left\{\omega : \alpha < X(\omega) \leq \beta - \frac{1}{n}\right\} \in \mathcal{F}$.

Thus, $\{X \in A\} \in \mathcal{F}$ is an event for both open and closed A. This technical definition of random variables also implies that they satisfy certain algebraic properties:

- If X, Y are random variables, then so are $X + Y$, XY and X^p for $p \in \mathbb{R}$
- If X is a random variable, then so is $f(X)$ for most functions f. However, there are counterexamples of "pathological" functions that are way beyond the scope of this class (non-Borel measurable functions).
- If $X_1, X_2, \ldots$ are random variables, then so is $\lim_{n \to \infty} X_n$ (if the limit exists)

3.2 Discrete Random Variables

Definition 3.2: Discrete Random Variables
A discrete random variable is one that takes countably many values.

Some basic examples are

- X = roll of a dice (takes values in $\{1, 2, 3, 4, 5, 6\}$)
- X = number of times I need to cast a fishing rod before I catch a fish
- X = fraction of heads in n coin flips (non-integer values are acceptable)

Definition 3.3: Probability Mass Function
The frequencies with which a discrete random variable takes different values is defined by its probability mass function (PMF)

$$P_X(x) = \mathbb{P}[X = x] = P(\{\omega : X(\omega) = x\})$$

The PMF is also called the distribution of X.

The probability mass function does not depend on the individual sample points $\omega \in \Omega$. It is also defined as $P_X : \mathcal{X} \to [0, 1]$ where $\mathcal{X}$ is the range of X.

Example 3.1
If X is a fair coin flip such that heads is 1 and tails is 0, then $P_X(1) = \frac{1}{2}$ and $P_X(0) = \frac{1}{2}$.

3.3 Joint Distributions

If X, Y are random variables on a common probability space $(\Omega, \mathcal{F}, P)$, then their "joint PMF" describes the frequencies of their joint outcomes.

$$\begin{aligned} P_{XY}(x,y) &= \mathbb{P}[X = x, Y = y] \\ &= P(\{\omega : X(\omega) = x, Y(\omega) = y\}) \\ &= P(\{\omega : X(\omega) = x\} \cap \{\omega : Y(\omega) = y\}) \end{aligned}$$

Some consequences of joint distributions are

- By the law of total probability,
$$\sum_y P_{XY}(x,y) = P_X(x)$$
This is called the marginal distribution of X.
- Random variables X, Y are independent iff $P_{XY}(x,y) = P_X(x)P_Y(y) \iff \{\omega : X(\omega) = x\}$ and $\{\omega : Y(\omega) = y\}$ are independent in $(\Omega, \mathcal{F}, P)$, $\forall x, y \in \mathcal{X} \times \mathcal{Y}$

Example 3.2
X_1, X_2 are fair coins that are magically linked together such that $X_1 = X_2$ with probability $\frac{3}{4}$. Then,

$$\begin{aligned} P_{XY}(0,0) = P_{XY}(1,1) &= \frac{3}{8} \\ P_{XY}(0,1) = P_{XY}(1,0) &= \frac{1}{8} \end{aligned}$$

Note 3.1
Often times, it is the most natural to model a problem just in terms of random variables and their (joint) distributions, and no mention is made regarding the probability space. This turns out to be ok! There is a deep theorem in probability called the Kolmogorov Extension Theorem which says if random variables and their distributions are specified in a "consistent" way, then there exists an underlying probability space that gives rise to the desired joint distributions.

Note 3.2
Explanation of notation: Pr (or $\mathbb{P}[]$) is used to denote a generic probability assignment when the probability space has not been explicitly defined.

3.4 Expectation

Definition 3.4: Expectation
For a discrete random variable X, the expectation of X is defined as

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x P_X(x)$$

provided the sum actually exists (i.e., it does not blow up to infinity).

Theorem 3.1: Law of The Unconscious Statistician
For $g : \mathcal{X} \to \mathbb{R}$, we have

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) P_X(x)$$

Proof: Note that $Y = g(X)$ is a random variable. Then,

$$\begin{aligned}\mathbb{E}[Y] &= \sum_y y P_Y(y) \\ &= \sum_y y \sum_{x:g(x)=y} P_X(x) \\ &= \sum_y \sum_{x:g(x)=y} g(x) P_X(x) \\ &= \sum_{x\in\mathcal{X}} g(x) P_X(x)\end{aligned}$$

This shows that you can compute $\mathbb{E}[Y]$ for $Y = g(X)$ without having to compute $P_Y(y)$.

The notion of expectation and LOTUS also extends to joint distributions. For RVs $X_1, \ldots, X_n$, we have

$$\mathbb{E}[f(X_1, \ldots, X_n)] = \sum_{x_1} \cdots \sum_{x_n} f(x_1, \ldots, x_n) P_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$$

Theorem 3.2: Linearity of Expectation
$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ for $a, b \in \mathbb{R}$

Proof:

$$\begin{aligned}\mathbb{E}[aX] &= \sum axP_X(x) \\ &= a\sum xP_X(x) \\ &= a\mathbb{E}[X] \\ \mathbb{E}[X+Y] &= \sum_{x,y}(x+y)P_{XY}(x,y) \\ &= \sum_x x\sum_y P_{XY}(x,y) + \sum_y y\sum_x P_{XY}(x,y) \\ &= \sum_x xP_X(x) + \sum_y yP_Y(y) \\ &= \mathbb{E}[X] + \mathbb{E}[Y]\end{aligned}$$

Linearity of Expectation works without any assumption of independence.

Example 3.3
Let X_1 and X_2 denote outcomes of rolling two fair dice.

$$\begin{aligned}\mathbb{E}[X_1 + X_2] &= \mathbb{E}[X_1] + \mathbb{E}[X_2] \\ &= \frac{7}{2} + \frac{7}{2} \\ &= 7\end{aligned}$$

Example 3.4
Let X_1 be the roll of a fair dice, and X_2 equal 7− (first roll).

$$\begin{aligned}\mathbb{E}[X_1 + X_2] &= \mathbb{E}[X_1] + \mathbb{E}[X_2] \\ &= \frac{7}{2} + \frac{7}{2}\end{aligned}$$

$$= 7$$

Definition 3.5: Indicator RV

An indicator for an event $A \in \mathcal{F}$ is defined as an RV such that

$$I = 1_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

Note 3.3

For an indicator I, note that

$$\begin{aligned} \mathbb{E}[I] &= 1 \cdot \mathbb{P}[I = 1] + 0 \cdot \mathbb{P}[I = 0] \\ &= \mathbb{P}[I = 1] \\ \mathbb{E}\left[I^2\right] &= 1^2 \cdot \mathbb{P}[I = 1] + 0^2 \cdot \mathbb{P}[I = 0] \\ &= \mathbb{P}[I = 1] \\ &= \mathbb{E}[I] \end{aligned}$$

Note 3.4

If I_A and I_B are indicators for two events A and B, then

$$\begin{aligned} \mathbb{E}[I_A I_B] &= \mathbb{P}[I_A I_B = 1] \\ &= \mathbb{P}[I_A = 1, I_B = 1] \\ &= P(A \cap B) \end{aligned}$$

Example 3.5

A really powerful technique for solving problems in probability is to introduce indicators and use the linearity of expectation. One famous example is that of derangements: suppose n people put their hats into a basket and draw one out at random. What is the expected number of people who get their own hat? The main idea is to introduce indicator random variables

$$X_i = \begin{cases} 1 & \text{if person } i \text{ gets their own hat} \\ 0 & \text{otherwise} \end{cases}$$

such that

$$\begin{aligned} \mathbb{E}[\text{people who get their own hat}] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}[X_i] \\ &= \sum_{i=1}^{n} \mathbb{P}[X_i = 1] \\ &= n\frac{1}{n} \\ &= 1 \end{aligned}$$

Theorem 3.3: Tail Sum Formula

For non-negative integer valued random variables, we have

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}[X \geq k]$$

Proof: The proof relies on the change of bounds given below

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k \geq 1} k P_X(k) \\
&= \sum_{k \geq 1} \sum_{j=1}^{k} P_X(k) \\
&= \sum_{j \geq 1} \sum_{k \geq j} P_X(k) \\
&= \sum_{j \geq 1} \mathbb{P}[X \geq j]
\end{aligned}$$

Example 3.6

Roll 4 fair die and let $M = \min(X_1, X_2, X_3, X_4)$.

$$\begin{aligned}
\mathbb{E}[M] &= \sum_{k \geq 1} \mathbb{P}[M \geq k] \\
&= \sum_{k \geq 1} \prod_{i=1}^{4} \mathbb{P}[X_i \geq k] \\
&= \sum_{k=1}^{6} \left(\frac{6 - k + 1}{6} \right)^4 \\
&\approx 1.75
\end{aligned}$$

4 Lecture 4

4.1 Variance

Definition 4.1: Variance
The variance of a random variable is defined as

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \sum_x (x - \mathbb{E}[X])^2 P_X(x) \\ &= \sum_x x^2 P_X(x) - 2\mathbb{E}[X] \sum_x x P_X(x) + \mathbb{E}[X]^2 \sum_x P_X(x) \\ &= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2\end{aligned}$$

In general, for random variable X and constants $a, b \in \mathbb{R}$,

$$\begin{aligned}\text{Var}(aX) &= \mathbb{E}\left[(aX)^2\right] - \mathbb{E}[aX]^2 \\ &= a^2\mathbb{E}\left[X^2\right] - a^2\mathbb{E}[X]^2 \\ &= a^2\left(\mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2\right) \\ &= a^2 \text{Var}(X) \\ \text{Var}(X+b) &= \mathbb{E}\left[(X + b - \mathbb{E}[X+b])^2\right] \\ &= \mathbb{E}\left[(X + b - \mathbb{E}[X] - b)^2\right] \\ &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \text{Var}(X)\end{aligned}$$

Definition 4.2: Standard Deviation
The standard deviation of X is defined as $\sigma_X = \sqrt{\text{Var}(X)}$. This is why $\text{Var}(X)$ is sometimes also denoted by σ_X^2.

Definition 4.3: Covariance
Covariance can be defined as a measure of dependence between X and Y.

$$\begin{aligned}\text{Cov}(X,Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

Theorem 4.1: Sum of Variances
If X, Y are independent, i.e., $P_{XY}(x,y) = P_X(x)P_Y(y)$, then $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$.

Proof: This is a multi-step proof:

- Step 1: If X, Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

$$\mathbb{E}[XY] = \sum_{x,y} xy P_{XY}(x,y)$$

$$
\begin{aligned}
&= \sum_{x,y} xy P_X(x) P_Y(y) \\
&= \left(\sum_x x P_X(x)\right)\left(\sum_y y P_Y(y)\right) \\
&= \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}
$$

- Step 2:

$$
\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}\left[(X+Y-\mathbb{E}[X+Y])^2\right] \\
&= \mathbb{E}\left[(X-\mathbb{E}[X]+Y-\mathbb{E}[Y])^2\right] \\
&= \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] + \mathbb{E}\left[(Y-\mathbb{E}[Y])^2\right] + 2\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)
\end{aligned}
$$

Since $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$ for independent RVs, we have that $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Note 4.1
$\mathrm{Var}(X) = \mathrm{Cov}(X,X)$

Note 4.2
Covariance is bilinear, i.e.,

$$\mathrm{Cov}\left(\sum_i \alpha_i X_i, \sum_j \beta_j Y_j\right) = \sum_i \sum_j \alpha_i \beta_j \,\mathrm{Cov}(X_i, Y_j)$$

Definition 4.4: Correlation Coefficient
The correlation coefficient of two RVs can be defined as

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Just like covariance, correlation can be interpreted as a measure of the dependence between two RVs. However, it is often seen as a slightly more meaningful metric since the normalization above bounds its value between -1 and 1.

Proof: The bounds follow from the Cauchy-Schwartz inequality. WLOG, let $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. Then,

$$
\begin{aligned}
|\mathbb{E}[XY]| &= \left|\sum_{x,y} xy P_{XY}(x,y)\right| \\
&= \left|\sum_{x,y} \left(x\sqrt{P_{XY}(x,y)}\right)\left(y\sqrt{P_{XY}(x,y)}\right)\right| \\
&\le \left|\sqrt{\left(\sum_{x,y} x^2 P_{XY}(x,y)\right)}\sqrt{\left(\sum_{x,y} y^2 P_{XY}(x,y)\right)}\right| \\
&= \left|\sqrt{\left(\sum_x x^2 P_X(x)\right)}\sqrt{\left(\sum_y y^2 P_Y(y)\right)}\right| \\
&= \left|\sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}\right|
\end{aligned}
$$

Thus, $\left|\frac{\mathbb{E}[XY]}{\sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}}\right| = |\rho(X,Y)| \le 1.$

The inner product of two random variables is defined as the expectation of their product. We will come back to this notion when we discuss the Hilbert space of random variables later in the course.

Example 4.1: Uncorrelated $\not\Rightarrow$ Independence
Consider the following joint distribution:

$$P_{XY}(-1,0) = P_{XY}(1,0) = P_{XY}(0,-1) = P_{XY}(0,1) = \frac{1}{4}$$

Then,

$$\mathbb{P}[X=0] = \mathbb{P}[X=0,Y=-1] + \mathbb{P}[X=0,Y=1] = \frac{1}{4}+\frac{1}{4} = \frac{1}{2}$$
$$\mathbb{P}[Y=0] = \mathbb{P}[Y=0,X=-1] + \mathbb{P}[Y=0,X=1] = \frac{1}{4}+\frac{1}{4} = \frac{1}{2}$$

but

$$\mathbb{P}[X=0,Y=0] = 0 \neq \frac{1}{2}\cdot\frac{1}{2} = \mathbb{P}[X=0]\mathbb{P}[Y=0]$$

Thus, X and Y are clearly not independent. However,

$$\mathbb{E}[XY] = (0)(-1)\frac{1}{4} + (0)(1)\frac{1}{4} + (-1)(0)\frac{1}{4} + (1)(0)\frac{1}{4} = 0$$
$$\mathbb{E}[X] = -1\cdot\frac{1}{4} + 0\cdot\frac{1}{2} + 1\cdot\frac{1}{4} = 0$$
$$\mathbb{E}[Y] = -1\cdot\frac{1}{4} + 0\cdot\frac{1}{2} + 1\cdot\frac{1}{4} = 0$$

making $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$.

4.2 Common Examples of Discrete Distributions

- $X \sim \mathrm{Uniform}(\{1,\dots,n\})$ has a PMF given by

$$P_X(k) = \begin{cases} \frac{1}{n} & \text{for } k \in \{1,\dots,n\} \\ 0 & \text{otherwise} \end{cases}$$
$$\mathbb{E}[X] = \frac{n}{2}$$
$$\mathrm{Var}(X) = \frac{n^2-1}{12}$$

- $X \sim \mathrm{Bernoulli}(p)$ has a PMF given by

$$P_X(k) = \begin{cases} 1-p & k=0 \\ p & k=1 \\ 0 & \text{otherwise} \end{cases}$$
$$\begin{aligned}\mathbb{E}[X] &= 0\cdot(1-p) + 1\cdot p \\ &= p\end{aligned}$$
$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\ &= 0\cdot(1-p) + 1\cdot p - p^2 \\ &= p(1-p)\end{aligned}$$

  One use for Bernoulli random variables is that they can model indicators i.e. $1_A \sim \mathrm{Bernoulli}(P(A))$.

- $X \sim \mathrm{Binomial}(n,p)$ has a PMF given by

$$P_X(k) = \begin{cases} \binom{n}{k}p^k(1-p)^{n-k} & k \in \{0,\dots,n\} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}[X] = np$$
$$\text{Var}(X) = np(1-p)$$

The expressions above were derived by letting $X = \sum_i X_i$ where $X_i \sim \text{Bernoulli}(p)$ are IID indicator variables.

- $X \sim \text{Geometric}(p)$ has a PMF given by

$$P_X(k) = \begin{cases} p(1-p)^{k-1} & k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$
$$\begin{aligned} \mathbb{P}[X > k] &= 1 - \mathbb{P}[X \leq k] \\ &= 1 - \sum_{i=0}^{k} p(1-p)^{i-1} \\ &= 1 - p\frac{1-(1-p)^k}{1-(1-p)} \\ &= (1-p)^k \end{aligned}$$
$$\begin{aligned} \mathbb{E}[X] &= \sum_{k \geq 1} \mathbb{P}[X \geq k] \\ &= \sum_{k \geq 0} \mathbb{P}[X > k] \\ &= \sum_{k \geq 0} (1-p)^{k-1} \\ &= \frac{1}{p} \end{aligned}$$
$$\text{Var}(X) = \frac{1-p}{p^2}$$

Note that $X \sim \text{Geometric}(p)$ can also be described as $X \sim \min\{k \geq 1 : X_k = 1\}$ where $X_k \sim_{\text{IID}} \text{Bernoulli}(k)$.

- $X \sim \text{Poisson}(\lambda)$ has a PMF given by

$$P_X(k) = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda} & k \geq 0 \\ 0 & \text{otherwise} \end{cases}$$
$$\mathbb{E}[X] = \lambda$$
$$\text{Var}(X) = \lambda$$

A Poisson random variable models the number of Bernoulli arrivals in a fixed time interval interval. The parameter λ is the rate at which these arrivals occur. Since we are modelling the probability of some event occurring a certain number of times (Poisson's Law of Rare Events), we can define the Poisson distribution as a limit of the binomial. We claim that $X \sim \text{Binomial}(n, p)$ implies $\mathbb{E}[X] = np = \lambda$ as $n \to \infty$. Then,

$$\begin{aligned} \mathbb{P}[X = k] &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n(n-1)\dots(n-k-1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &\approx \frac{\lambda^k}{k!} e^{-\lambda} \end{aligned}$$

as $n \to \infty$

Note 4.3

By the construction of the Binomial distribution, if $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Binomial}(m, p)$ are independent, then $X + Y \sim \text{Binomial}(n + m, p)$. This follows since X and Y can be viewed as the number of successes from m and n

IID Bernoulli trials respectively, making $X+Y$ the number of successes collectively from $m+n$ total Bernoulli trials. Similarly, if $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ are independent, then $X+Y \sim \text{Poisson}(\lambda+\mu)$.

Example 4.2: Coupon Collector's Problem

Suppose I buy boxes of cereal, and each box contains a random coupon (out of N total possible). How many boxes of cereal do I need to buy until I collect all N coupons?

Let X_i be the number of boxes that I need to buy to get the ith coupon, starting from when I found the $(i-1)$th coupon. If X is the number of boxes I need to buy, then $X = \sum_{i=1}^{N} X_i$. Also,

$$X_i \sim \text{Geometric}\left(1 - \frac{i-1}{N}\right) = \text{Geometric}\left(\frac{N-i+1}{N}\right)$$

because we only have $N-(i-1)$ possible new coupons left to collect after already getting $i-1$ of them. Applying linearity of expectation,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{N} \mathbb{E}[X_i] \\
&= \frac{N}{N} + \frac{N}{N-1} + \frac{N}{N-2} + \cdots + \frac{N}{1} \\
&= N\left(1 + \frac{1}{2} + \cdots + \frac{1}{N}\right) \\
&\approx N \log N
\end{aligned}$$

4.3 Conditional Distributions

Definition 4.5: Conditional PMF

The conditional PMF of an RV X, conditioned on the event A, is defined as

$$\begin{aligned}
P_{X|A}(x) &= \mathbb{P}[X = x \mid A] \\
&= \frac{\mathbb{P}[X = x \cap A]}{\mathbb{P}[A]}
\end{aligned}$$

For discrete random variables, we can define the conditional PMF of X conditioned on $Y = y$, provided $P_Y(y) > 0$, as

$$\begin{aligned}
P_{X|Y}(x \mid y) &= \mathbb{P}[X = x \mid Y = y] \\
&= \frac{\mathbb{P}[X = x \cap Y = y]}{\mathbb{P}[Y = y]} \\
&= \frac{P_{XY}(x, y)}{P_Y(y)}
\end{aligned}$$

The conditional PMF of X can be interpreted as the new distribution of X given some event A or $\{Y = y\}$ occurs.

Example 4.3

We pick up coin 1 $\left(\text{bias } = \frac{1}{4}\right)$ with probability $\frac{1}{2}$ and coin 2 $\left(\text{bias } = \frac{3}{4}\right)$ with probability $\frac{1}{2}$. Toss the chosen coin twice and let $H = 1$ and $T = 0$. Let Y = first toss and X = second toss. Then,

$$\begin{aligned}
P_{X|Y}(1 \mid 1) &= \frac{P_{XY}(1,1)}{P_Y(1)} \\
&= \frac{\frac{1}{2}\left(\frac{1}{4}\right)^2 + \frac{1}{2}\left(\frac{3}{4}\right)^2}{\frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{3}{4}}
\end{aligned}$$

$$= \frac{5}{8}$$

Example 4.4: Memoryless Property

If $X \sim \text{Geometric}(p)$, then,

$$\mathbb{P}[X = k + m \mid X > k] = \mathbb{P}[X = m]$$

This is the same as $P_{X|Y}(k + m \mid 1)$ where $Y = 1_{\{X>k\}}$.

Example 4.5

Let X and Y be two random variables. Since $A = \{X = x\}$ and $B = \{Y = y\}$ are events, by Bayes Rule

$$\begin{aligned}\mathbb{P}[X = x \mid Y = y] &= \mathbb{P}[A \mid B] \\ &= \frac{\mathbb{P}[B \mid A]\mathbb{P}[A]}{\mathbb{P}[B]} \\ &= \frac{\mathbb{P}[Y = y \mid X = x]\mathbb{P}[X = x]}{\mathbb{P}[Y = y]}\end{aligned}$$

Example 4.6

Let X be the roll of a dice and A be the event that it is an even number. Then,

$$P_{X|A}(x) = \frac{\mathbb{P}[X = x \cap x \text{ is even}]}{\mathbb{P}[A]}$$

$$\mathbb{P}[A] = \frac{3}{6} = \frac{1}{2}$$

$$\mathbb{P}[X = x \cap x \text{ is even}] = \begin{cases}\frac{1}{6} & x = 2, 4, 6 \\ 0 & x = 1, 3, 5\end{cases}$$

$$P_{X|A}(x) = \begin{cases}\frac{1}{3} & x = 2, 4, 6 \\ 0 & x = 1, 3, 5\end{cases}$$

Note 4.4

Recall that two events A and B are independent if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$ or $\mathbb{P}[A \mid B] = \mathbb{P}[A]$. This notion can be extended even further: a random variable X is independent of event A iff

$$\mathbb{P}[X = x \cap A] = P_X(x)\mathbb{P}[A]$$

$$P_{X|A}(x) = P_X(x)$$

5 Lecture 5

5.1 Conditional Expectation

Definition 5.1: Conditional Expectation

The expected value of X given that I know $Y = y$ is

$$\mathbb{E}[X \mid Y = y] = \sum_x x P_{X|Y}(x \mid y)$$

This can be generalized for any event A such that $\mathbb{P}[A] > 0$:

$$\mathbb{E}[X \mid A] = \sum_x x P_{X|A}(x)$$

Note 5.1

$\mathbb{E}[X \mid Y = y]$ is a function of y — the conditional expectation can change for different values of y

As we saw before, if Z is a random variable, then so is $g(Z)$ for some function g. Then, $g(Y) = \mathbb{E}[X \mid Y]$ is also a random variable which, again, is a function of Y: it takes on the value $g(y) = \mathbb{E}[X \mid Y = y]$ with probability $\mathbb{P}[Y = y]$.

Theorem 5.1: Tower Property

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

for all functions f.

Proof:

$$\begin{aligned}
\mathbb{E}[f(Y)X] &= \sum_{x,y} f(y) x P_{XY}(x, y) \\
&= \sum_{x,y} f(y) x P_{X|Y}(x \mid y) P_Y(y) \\
&= \sum_y f(y) \left(\sum_x x P_{X|Y}(x \mid y) \right) P_Y(y) \\
&= \sum_y f(y) \mathbb{E}[X \mid Y = y] P_Y(y) \\
&= \mathbb{E}[f(Y)\mathbb{E}[X \mid Y]]
\end{aligned}$$

Theorem 5.2: Iterated Expectation

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

for $f(Y) = 1$ and using the tower property.

Theorem 5.3: Total Expectation

Expanding out the statement above, we get the law of total expectation,

$$\mathbb{E}[X] = \sum_y \mathbb{E}[X \mid Y = y] \mathbb{P}[Y = y]$$

Again, this can be further generalized to disjoint events A_i that partition Ω with $\mathbb{P}[A_i] > 0$ such that

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X \mid A_i] \mathbb{P}[A_i]$$

Example 5.1
Toss a fair coin N times, and let H be the number of heads. Then,

$$\begin{aligned}\mathbb{E}[H] &= \mathbb{E}[\mathbb{E}[H \mid N]] \\ &= \mathbb{E}\left[\frac{N}{2}\right] \\ &= \frac{1}{2}\mathbb{E}[N]\end{aligned}$$

5.2 Continuous Random Variables

Recall that discrete random variables are defined by a PMF

$$\mathbb{P}[X \in B] = \sum_{X \in B} P_X(x)$$

Definition 5.2: Probability Density Function
For a continuous random variable, the distribution is defined via its "density" $f_X : \mathbb{R} \to [0, \infty]$. That is,

$$\mathbb{P}[X \in B] = \int_B f_X(x)\,\mathrm{d}x$$

Densities must satisfy

1. $f_X \geq 0$
2. $\int_{\mathbb{R}} f_X(x)\,\mathrm{d}x = 1$

Observe that for a continuous random variable, $\mathbb{P}[X = x] = 0$ since

$$\begin{aligned}\mathbb{P}[X = x] &\leq \mathbb{P}[x \leq X < x + \delta] \\ &= \int_x^{x+\delta} f_X(u)\,\mathrm{d}u \\ &\approx \delta f_X(x)\end{aligned}$$

As $\delta \to \infty \implies \mathbb{P}[X = x] \to 0$. Equivalently,

$$f_X(X) = \lim_{\delta \to 0} \frac{\mathbb{P}[X \in [x, x + \delta]]}{\delta}$$

Definition 5.3: Cumulative Density Function
The cumulative distribution function (CDF) of a random variable X is defined as

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

A CDF must satisfy these properties in general:

1. $\lim_{x \to -\infty} F_X(x) = 0$ because $\mathbb{P}[X \leq -\infty] = 0$
2. $\lim_{x \to -\infty} F_X(x) = 1$ because $\mathbb{P}[X \leq \infty] = 1$
3. $\lim_{y \to x^+} F_X(y) = F_X(x)$, also known as right continuity

These properties follow directly from the Kolmogorov axioms and the definition of random variables.

For a continuous random variable, this is

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u$$

Observe that

$$\begin{aligned}\mathbb{P}[a \le X \le b] &= \int_a^b f_X(x)\,\mathrm{d}x \\ &= \int_{-\infty}^b f_X(x)\,\mathrm{d}x - \int_{-\infty}^a f_X(x)\,\mathrm{d}x \\ &= F_X(a) - F_X(b)\end{aligned}$$

Then, by the fundamental theorem of calculus, $f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x}F_X(x)$.

5.2.1 Extensions to multiple random variables

We say $X_1, X_2, \dots X_n$ are (jointly) continuous random variables if there is a function

$$f_{X_1,X_2,\dots,X_n} : \mathbb{R}^n \to [0,\infty]$$

such that

$$\begin{aligned}F_{X_1,X_2,\dots,X_n}(x_1,x_2,\dots,x_n) &= \mathbb{P}[X_1 \le x_1, X_2 \le x_2, \dots, X_n \le x_n] \\ &= \int_{-\infty}^{x_n}\cdots\int_{-\infty}^{x_1} f_{X_1,\dots,X_n}(u_1,\dots,u_n)\,\mathrm{d}u_1\dots\mathrm{d}u_n\end{aligned}$$

and

$$\int_{\mathbb{R}}\cdots\int_{\mathbb{R}} f_{X_1,\dots,X_n}(x_1,\dots,x_n)\,\mathrm{d}x_1\dots\mathrm{d}x_n = 1$$

Just like marginal distributions for discrete RVs, we can define the concept of marginal density for continuous RVs as well. If X, Y are jointly continuous with joint density $f_{XY}(x,y)$, then

$$\begin{aligned}f_X(x) &= \int_{\mathbb{R}} f_{XY}(x,y)\,\mathrm{d}y \\ f_Y(y) &= \int_{\mathbb{R}} f_{XY}(x,y)\,\mathrm{d}x\end{aligned}$$

5.3 Independence

Definition 5.4: Independence of continuous RVs
If X, Y are continuous, they are independent iff

$$f_{XY}(x,y) = f_X(x)f_Y(y)$$

Note 5.2
Two random variables (both discrete and continuous) are independent iff

$$F_{XY}(x,y) = F_X(x)F_Y(y)$$

Example 5.2
If we throw a dart at a dartboard of radius r, let (X,Y) denote the pair of x-y coordinate of where it lands. If a dart lands uniformly on the board, then,

$$f_{XY}(x,y) = \begin{cases}\frac{1}{\pi r^2} & x^2+y^2 \le r^2 \\ 0 & \text{otherwise}\end{cases}$$

Example 5.3
The (X, Y) in the dartboard example above are not independent.

Example 5.4
If (X, Y) is uniform on the box $[0, r] \times [0, r]$, then X, Y are independent since

$$\begin{aligned} f_{XY}(x, y) &= \frac{1}{r^2} 1_{\{X \in [0,r], Y \in [0,r]\}} \\ &= \frac{1}{r} 1_{\{x \in [0,r]\}} \frac{1}{r} 1_{\{y \in [0,r]\}} \\ &= f_X(x) f_Y(y) \end{aligned}$$

5.4 Common Examples of Continuous Random Variables

1. $X \sim \text{Uniform}(a, b)$

$$\begin{aligned} f_X(x) &= \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases} \\ F_X(x) &= \int_{-\infty}^{x} f_X(u)\, \mathrm{d}u \\ &= \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & a \ge b \end{cases} \end{aligned}$$

2. $X \sim \text{Exp}(\lambda)$

$$\begin{aligned} f_X(x) &= \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases} \\ F_X(x) &= \lambda \int_{-\infty}^{x} e^{-\lambda u}\, \mathrm{d}u \\ &= 1 - e^{-\lambda x} \end{aligned}$$

5.5 Expectation and Variance of Continuous Random Variables

Expectation for continuous random variables is similar to the discrete case:

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x)\, \mathrm{d}X$$

More generally,

$$\begin{aligned} \mathbb{E}[g(X)] &= \int_{\mathbb{R}} g(x) f_X(x)\, \mathrm{d}x \\ \mathbb{E}[g(X_1, \cdots, X_n)] &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} g(x_1, \ldots, x_n) f_{X_1, \ldots, X_n}(x_1, \ldots, x_n)\, \mathrm{d}x_1 \ldots \mathrm{d}x_n \end{aligned}$$

Variance and covariance are defined the same way for both discrete and continuous random variables.

Example 5.5

$X \sim \text{Uniform}(a, b)$. Then,

$$\begin{aligned}
\mathbb{E}[X] &= \frac{1}{b-a}\int_a^b x\,\mathrm{d}x \\
&= \frac{a+b}{2} \\
\mathbb{E}\left[X^2\right] &= \frac{1}{b-a}\int_a^b x^2\,\mathrm{d}x \\
&= \frac{1}{3}\frac{b^3-a^3}{b-a} \\
\text{Var}(X) &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\
&= \frac{(b-a)^2}{12}
\end{aligned}$$

Example 5.6

$X \sim \text{Exp}(\lambda)$. Then,

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty \lambda x e^{-\lambda x}\,\mathrm{d}x \\
&= \frac{1}{\lambda} \\
\mathbb{E}\left[X^2\right] &= \int_0^\infty \lambda x^2 e^{-\lambda x}\,\mathrm{d}x \\
&= \frac{2}{\lambda^2} \\
\text{Var}(X) &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\
&= \frac{1}{\lambda^2}
\end{aligned}$$

Example 5.7

Back to the dartboard example - Let R = distance from dart to center = $\sqrt{X^2+Y^2}$.

$$\begin{aligned}
\mathbb{P}\left[R \le \frac{r}{2}\right] &= \mathbb{P}\left[X^2+Y^2 \le \frac{r^2}{4}\right] \\
&= \mathbb{E}\left[1_{\left\{X^2+Y^2 \le \frac{r^2}{4}\right\}}\right] \\
&= \frac{1}{\pi r^2}\iint 1_{\left\{x^2+y^2 \le \frac{r^2}{4}\right\}}\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{\pi r^2}\frac{\pi r^2}{4} \\
&= \frac{1}{4}
\end{aligned}$$

5.6 Conditioning

Definition 5.5: Conditional Density

Let X, Y be continuous random variables. The conditional density of X given $Y = y$ is

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

Similarly, when conditioning on an event instead of a random variable,

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{\mathbb{P}[A]} & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$
$$= \begin{cases} \frac{\int f_{XY}(x,y)\,\mathrm{d}y}{\mathbb{P}[A]} & (x,y) \in A \\ 0 & \text{otherwise} \end{cases}$$

Note 5.3
The definitions above are only a device to simplify calculations. They are not a true "conditional probability" in the sense that $\mathbb{P}[Y = y] = 0$.

With conditional density, we can define conditional expectation for continuous random variables as

$$\mathbb{E}[X \mid Y = y] = \int x f_{X|Y}(x \mid y)\,\mathrm{d}x$$

Let $\mathbb{E}[X \mid Y]$ denote the function above evaluated at Y. The tower property will still hold:

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

The laws of iterated and total expectation can also be calculated as

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$$
$$= \int \mathbb{E}[X \mid Y = y] f_Y(y)\,\mathrm{d}y$$

In general, you can *usually* replace the summation in expressions for discrete random variables with integrals, and the PMFs with PDFs whenever appropriate.

6 Lecture 6

6.1 Review

Let $(\Omega, \mathcal{F}, P)$ be a probability space that defines our universe. Moreover, let $X : \Omega \to \mathcal{R}$ be a random variable. Then, the "distribution of X" refers to the frequency at which X takes given values. It is captured by the cumulative distribution function

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

for $x \in \mathbb{R}$. It satisfies three important properties:

- F_X is non-decreasing
- $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{X \to +\infty} F_X(x) = 1$
- F_X is right continuous

Note 6.1
Any F satisfying these properties is the CDF of some random variable in some probability space

- If X is discrete, then
$$F_X(x) = \sum_{x' \leq x} P_X(x)$$
for some function $P_X : \mathcal{X} \to [0, 1]$ called the PMF of X. In this case, F_X is a staircase function.
- If X is continuous, then
$$F_X(x) = \int_{-\infty}^{x} f_X(x)\, \mathrm{d}x$$
for some function f_X called the PDF of X.
- Extension to multiple random variables: if X and Y are discrete, then P_{XY} is their joint PMF but if they are continuous, then f_{XY} is their joint density.

With these, we can compute things like expectation, variance, conditional PMF/PDF etc.

6.2 Memoryless Property

Suppose we want a continuous random variable X with a "memoryless" property like that of a geometric random variable. Mathematically, we want a random variable X satisfying

$$\begin{aligned}
\mathbb{P}[X > t + s \mid X > s] &= \mathbb{P}[X > t], \forall s, t \geq 0 \\
\frac{\mathbb{P}[X > t + s \cap X > s]}{\mathbb{P}[X > s]} &= \mathbb{P}[X > t], \forall s, t \geq 0 \\
\mathbb{P}[X > t + s] &= \mathbb{P}[X > t]\mathbb{P}[X > s]
\end{aligned}$$

Let $\mathbb{P}[X > x] = f(x)$. Then, we want that

$$f(t + s) = f(t) \cdot f(s)$$

Only non-zero solution to this functional equation is $f(x) = e^{\alpha x}$ for some $\alpha \in \mathbb{R}$. Then,

$$\begin{aligned}
F_X(t) &= \mathbb{P}[X \leq t] \\
&= 1 - \mathbb{P}[X > t] \\
&= 1 - e^{\alpha t} \text{ for some } \alpha < 0
\end{aligned}$$

Now, we can conclude that the only random variables with this property have a CDF of the form

$$F_X(t) = \begin{cases} 1 - e^{-\lambda t} & \text{for } \lambda > 0, t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

This is precisely the CDF of an exponential random variable, meaning that it is the only continuous distribution with the memoryless property.

Intuitively, the memoryless property dictates that if a system is following a memoryless RV, its history will not influence its future. Say you are tossing a coin until a heads shows up. Since all tosses are independent, this can be modeled using a geometric distribution. From your perspective, the probability that you toss at least t times, starting at a given moment, is the same regardless of whether you have already tossed 0 or s coins initially. In other words, does it really matter what you did with the coin before you start your tossing experiment? No! If it did, that would imply the existence of a "max number of tosses", but that defies logic since you can theoretically land on tails infinitely many times (although improbable, it is not impossible)!

You can view an exponential distribution as the limit of a geometric RV, where a single Bernoulli trials occurs over an infinitesimally small interval (similar to how a Poisson distribution is the limit of a binomial RV). Therefore, this analogy can also be extended to continuous RVs using exponential timers - the probability that you wait for at least a certain amount of time before a random timer goes off is the same regardless of the time elapsed so far. Exponential random variables tend to pop up naturally if the memoryless property is present in a certain situation. Some real life examples would include radioactive decay or random failures of different machine parts.

6.3 Gaussian Random Variable

Like exponential random variables, Gaussians emerge naturally in many ways, and so are another important class of RVs. There are many important applications of a gaussian distribution, some of which include

1. Central Limit Theorem (sum of many small independent effects is gaussian)
2. Modeling thermal noise in a resistor
3. Boltzmann model of interacting particles

Definition 6.1: Gaussian Distribution

X is a gaussian with mean μ and variance σ^2, denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$, if it has a PDF given by

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

For $\mu = 0$ and $\sigma^2 = 1$, we have $X \sim \mathcal{N}(0, 1)$ [standard normal] such that its CDF is defined as

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u$$

Unfortunately, there is no closed form for Φ. However, observe that

$$\begin{aligned} \Phi(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u - \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u \\ &= 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-x} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u \\ &= 1 - \Phi(-x) \end{aligned}$$

The change of bounds in line 3 follows from the evenness of the standard gaussian PDF. Thus, $\mathbb{P}[X \in [-x, x]] = \Phi(x) - \Phi(-x) = 2\Phi(x) - 1$.

Note 6.2
If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$. Conversely, if $X \sim \mathcal{N}(0,1)$, then $\sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2)$.

Proof: Proof that f_X (for a standard normal) is actually a density:

$$\begin{aligned}
\left(\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x\right)^2 &= \left(\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} f_X(y)\,\mathrm{d}y\right) \\
&= \iint_{\mathbb{R}^2} f_X(x) f_X(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{2\pi}\iint_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2}}\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2}} r\,\mathrm{d}r\,\mathrm{d}\theta \\
&= \int_0^{\infty} e^{-\frac{r^2}{2}}\,\mathrm{d}r \\
&= \int_0^{\infty} e^{-u}\,\mathrm{d}u \\
&= 1
\end{aligned}$$

Here are some cool facts about gaussians:

1. If $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ are independent, then $X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$. Be cautious that independence is needed here. Here is a counterexample for when X and Y are not independent

$$Y = \begin{cases} -X & |X| \le 1 \\ X & |X| > 1 \end{cases}$$

$$X + Y = \begin{cases} 0 & |X| \le 1 \\ 2X & |X| > 1 \end{cases}$$

   Both X and Y are gaussian here, but $X + Y$ clearly is not.

2. If X, Y are independent and $X + Y \sim \mathcal{N}(\mu, \sigma^2)$, then both X and Y are gaussian.

3. If X, Y are independent and $(X + Y), (X - Y)$ are independent, then both X and Y are gaussian.

Example 6.1
Let T = temperature of a chip on a satellite. Assume that $T \sim \mathcal{N}(20, 2)$. A failure is defined as the event that the temperature goes below $10°$ or above $50°$ in a 1 sec interval. Then,

$$\begin{aligned}
\mathbb{P}[\text{failure}] &= \mathbb{P}[X < 10] + \mathbb{P}[X > 50] \\
&= \mathbb{P}\left[\frac{T-20}{\sqrt{2}} < \frac{10-20}{\sqrt{2}}\right] + \mathbb{P}\left[\frac{T-20}{\sqrt{2}} > \frac{50-10}{\sqrt{2}}\right] \\
&= \Phi\left(\frac{-10}{\sqrt{2}}\right) + \left(1 - \Phi\left(\frac{30}{\sqrt{2}}\right)\right) \\
&= 7.7 \cdot 10^{-13}
\end{aligned}$$

Then, applying a union bound,

$$\begin{aligned}
\mathbb{P}[\text{failure after 25 years of spaceflight}] &\le (\text{num seconds in 25 years})\mathbb{P}[\text{failure}] \\
&\le (7.9 \cdot 10^8)(7.7 \cdot 10^{-13}) \\
&\le \frac{1}{1000}
\end{aligned}$$

Example 6.2

A cell phone send a signal $B \in \{-1, 1\}$ (equally likely) to a tower. The tower receives $Y = B + N$ where $N \sim \mathcal{N}(0, 1)$ is some thermal noise that is independent of B. Then,

$$f_{Y|B}(y \mid b) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-b)^2}{2}\right)$$

for $y \in \mathbb{R}$ and $b \in \{-1, 1\}$. Given that I observe $Y = y$, what is the probability that $B = +1$ was sent? By Bayes rule,

$$\begin{aligned} P_{B|Y}(+1 \mid y) &= \frac{P_B(+1)}{f_Y(y)} f_{Y|B}(y \mid +1) \\ &= \frac{\frac{1}{2}\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-1)^2}{2}\right)}{\frac{1}{2}\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y+1)^2}{2}\right) + \frac{1}{2}\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-1)^2}{2}\right)} \\ &= \frac{1}{1+e^{-2y}} \end{aligned}$$

7 Lecture 7

7.1 Conditional Variance

Recall that the conditional expectation $\mathbb{E}[X \mid Y = y]$ is the expectation of X with respect to $P_{X|Y}(\cdot \mid y)$ or $f_{X|Y}(\cdot \mid y)$. Conditional variance is defined similarly:

Definition 7.1: Conditional Variance

$$\begin{aligned}\mathrm{Var}(X \mid Y = y) &= \mathbb{E}\Big[(X - \mathbb{E}[X \mid Y = y])^2 \mid Y = y\Big] \\ &= \mathbb{E}\Big[X^2 \mid Y = y\Big] - (\mathbb{E}[X \mid Y = y])^2\end{aligned}$$

Just like conditional expectation $\mathbb{E}[X \mid Y]$ is the function $\mathbb{E}[X \mid Y = \cdot]$ evaluated at Y, the conditional variance $\mathrm{Var}(X \mid Y)$ is the function $\mathrm{Var}(X \mid Y = \cdot)$ evaluated at Y.

Example 7.1

$\mathbb{E}[\mathrm{Var}(X \mid Y)] = \mathbb{E}\big[(X - \mathbb{E}[X \mid Y])^2\big]$ is known as the minimum mean-square error of X given Y. It can be interpreted as how much "uncertainty" about X is reduced knowing Y (on average).

Theorem 7.1: Law of Total Variance

$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y])$

Proof:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\big[X^2\big] - \mathbb{E}[X]^2 \\ &= \mathbb{E}\big[\mathbb{E}\big[X^2 \mid Y\big]\big] - (\mathbb{E}[\mathbb{E}[X \mid Y]])^2 \\ &= \mathbb{E}\big[\mathrm{Var}(X \mid Y) + (\mathbb{E}[X \mid Y])^2\big] - (\mathbb{E}[\mathbb{E}[X \mid Y]])^2 \\ &= \mathbb{E}[\mathrm{Var}(X \mid Y)] + \Big(\mathbb{E}\big[(\mathbb{E}[X \mid Y])^2\big] - (\mathbb{E}[\mathbb{E}[X \mid Y]])^2\Big) \\ &= \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y])\end{aligned}$$

Note 7.1

The proof works entirely by manipulating things at the level of variance of expectation. Never needed to work at level of PMFs and PDFs

Example 7.2

Let $Y = X_1 + X_2 + \cdots + X_N$ where X_i are IID and N is a random variable taking values in $\{1, 2, \dots\}$.

$$\begin{aligned}\mathrm{Var}(Y) &= \mathbb{E}[\mathrm{Var}(Y \mid N)] + \mathrm{Var}(\mathbb{E}[Y \mid N]) \\ &= \mathbb{E}[N \,\mathrm{Var}(X_i)] + \mathrm{Var}(N\mathbb{E}[X_i]) \\ &= \mathrm{Var}(X_i)\mathbb{E}[N] + (\mathbb{E}[X_i])^2 \,\mathrm{Var}(N)\end{aligned}$$

7.2 Derived Distributions

The distribution of a random variable X is described by its CDF F_X. What if we have a new random variable $Y = g(X)$ for some function g. Howe do we find the distribution of Y?

First ask yourself if you really need the distribution of Y? Often times, we don't. For example,

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

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

However, if we really need it, then notice that

$$\begin{aligned} F_Y(y) &= \mathbb{P}[Y \le y] \\ &= \mathbb{P}[g(X) \le y] \\ &= \mathbb{P}[X \in \{x : g(x) \le y\}] \end{aligned}$$

In the case of a discrete random variables,

$$P_Y(y) = \sum_{x:g(x)=y} P_X(x)$$

If g is invertible, then

$$P_Y(y) = P_X(g^{-1}(y))$$

The formula above only works for discrete random variables.

Example 7.3

Let X be a continuous RV with the following PDF:

$$f_X(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Let $Y = 2X$. Blind application of discrete formula suggests

$$\begin{aligned} f_Y(y) &= f_X\left(\frac{y}{2}\right) \\ &= \begin{cases} 1 & 0 \le y \le 2 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

However, the PDF above does not integrate to 1 so this is a problem! For continuous random variables, it is better to work with their CDFs as stated above

$$\begin{aligned} F_Y(y) &= \mathbb{P}[Y \le y] \\ &= \mathbb{P}[2X \le y] \\ &= \mathbb{P}\left[X \le \frac{y}{2}\right] \\ &= F_X\left(\frac{y}{2}\right) \\ f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) \\ &= \frac{\mathrm{d}}{\mathrm{d}y} F_X\left(\frac{y}{2}\right) \\ &= \frac{1}{2} F_X'\left(\frac{y}{2}\right) \\ &= \frac{1}{2} f_X\left(\frac{y}{2}\right) \\ &= \begin{cases} \frac{1}{2} & 0 \le y \le 2 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

In general, for invertible function $g(x)$ such that $Y = g(X)$,

$$\begin{aligned} f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} \mathbb{P}[g(X) \le y] \\ &= \frac{\mathrm{d}}{\mathrm{d}y} \mathbb{P}\left[X \le g^{-1}(y)\right] \end{aligned}$$

$$= \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y))$$
$$= f_X(g^{-1}(y)) \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)$$
$$= f_X(g^{-1}(y)) \frac{1}{|g'(g^{-1}(y))|}$$

The absolute value accounts for monotonically decreasing functions g since their derivative would be negative, but the PDF of an RV is always non-negative. For linear functions like $Y = aX + b$, we have

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

7.3 Order Statistics

Let $X_1, X_2, \ldots, X_n \sim_{\text{IID}} X$, and sort them such that

$$X^{(1)} \leq X^{(2)} \leq \cdots \leq X^{(n)}$$

Then, $X^{(i)}$ is the ith smallest number in the list. That is, $X^{(1)}$ is $\min(X_1, \ldots, X_n)$ and $X^{(n)}$ is $\max(X_1, \ldots, X_n)$. What is the density of $X^{(i)}$?

$$f_{X^{(i)}}(y) = n\binom{n-1}{i-1} F_X(y)^{i-1} (1 - F_X(y))^{n-i} f_X(y)$$

Proof:

$$f_{X^{(i)}}(y)\,\mathrm{d}y \approx \mathbb{P}\left[X^{(i)} \in (y, y + \mathrm{d}y)\right]$$
$$\approx n\binom{n-1}{i-1} F_X(y)^{i-1} (1 - F_X(y))^{n-i} f_X(y)\,\mathrm{d}y$$

where

- $f_X(y)\,\mathrm{d}y$ is the probability that one of the $X^{(i)}$'s is in the interval $[y, y + \mathrm{d}y]$
- $F_X(y)^{i-1}$ is the probability that $i-1$ $X^{(i)}$'s are $\leq y$
- $(1 - F_X(y))^{n-i}$ is the probability that $n-i$ $X^{(i)}$'s are $\geq y + \mathrm{d}y$
- n ways to select the $X^{(i)}$ that falls in $[y, y + \mathrm{d}y]$
- $\binom{n-1}{i-1}$ number of ways to choose $i-1$ of the remaining $n-1$ $X^{(i)}$'s to be $\leq y$

Example 7.4
Suppose I model the time of bus k arriving as $X_k \sim \text{Exp}(\lambda)$ (all independent). Then, the arrival time of the ith soonest bus is given by $X^{(i)}$, whose density is

$$f_{X^{(i)}}(t) = n\binom{n-1}{i-1} \left(1 - e^{-\lambda t}\right)^{i-1} e^{-(n-i)\lambda t} \lambda e^{-\lambda t}$$

Example 7.5
The following two special cases pop up pretty frequently:

- If $Y = \max(X_1, \ldots, X_n)$ for $X_i \sim_{\text{IID}} X$, then

$$F_Y(y) = \mathbb{P}[Y \leq y]$$

$$\begin{aligned}
&= \mathbb{P}[\max(X_1, \dots, X_n) \le y] \\
&= \mathbb{P}[X_1 \le y] \dots \mathbb{P}[X_n \le y] \\
&= \mathbb{P}[X \le y]^n \\
&= F_X(y)^n
\end{aligned}$$

- If $Z = \min(X_1, \dots, X_n)$ for $X_i \sim_{\text{IID}} X$, then

$$\begin{aligned}
F_Z(z) &= \mathbb{P}[Z \le z] \\
&= \mathbb{P}[\min(X_1, \dots, X_n) \le z] \\
&= 1 - \mathbb{P}[\min(X_1, \dots, X_n) \ge z] \\
&= 1 - \mathbb{P}[X_1 \ge z] \dots \mathbb{P}[X_n \ge z] \\
&= 1 - (1 - \mathbb{P}[X_1 \le z]) \dots (1 - \mathbb{P}[X_n \le z]) \\
&= 1 - (1 - \mathbb{P}[X \le z])^n \\
&= 1 - (1 - F_X(z))^n
\end{aligned}$$

7.4 Convolutions

If X and Y are discrete, integer-valued random variables that are independent and $Z = X + Y$, what is the PMF of Z?

$$\begin{aligned}
P_Z(z) &= \mathbb{P}[X + Y = z] \\
&= \mathbb{P}\left[\bigcup_{k \in \mathbb{Z}} \{X = k\} \cap \{Y = z - k\}\right] \\
&= \sum_{k \in \mathbb{Z}} \mathbb{P}[X = k, Y = z - k] \\
&= \sum_{k \in \mathbb{Z}} P_X(k) P_Y(z - k) \\
&= (P_X * P_Y)(z)
\end{aligned}$$

Similarly, if X and Y are continuous with distributions f_X and f_Y instead, then $Z = X + Y$ has a density given by

$$\begin{aligned}
f_Z(z) &= \int_{\mathbb{R}} f_X(x) f_Y(z - x) \, \mathrm{d}x \\
&= (f_X * f_Y)(z)
\end{aligned}$$

Adding independent random variables corresponds to the convolution of their distributions.

Example 7.6

Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$. Then, letting $Z = X + Y$,

$$\begin{aligned}
P_Z(z) &= \sum_{k=0}^{z} P_X(k) P_Y(z - k) \\
&= \sum_{k=0}^{z} \frac{\lambda^k}{k!} e^{-\lambda} \frac{\mu^{(z-k)}}{(z-k)!} e^{-\mu} \\
&= \frac{1}{z!} e^{-(\lambda+\mu)} \sum_{k=0}^{z} \binom{z}{k} \lambda^k \mu^{z-k} \\
&= \frac{(\lambda + \mu)^z}{z!} e^{-(\lambda+\mu)}
\end{aligned}$$

Thus, $Z \sim \text{Poisson}(\lambda + \mu)$, which agrees with our intuition from section 4.2.

Example 7.7

Let $X, Y \sim \text{Uniform}(0,1)$. Then, letting $Z = X + Y$,

$$f_Z(z) = \int_{\mathbb{R}} f_X(x) f_Y(z-x)\,\mathrm{d}x$$

Note that $f_X(x) = 1$ for $0 \le x \le 1$ and $f_Y(z-x) = 1$ for $0 \le z - x \le 1$. We are only interested in the region bounded by those constraints since the integrand, and hence the density of Z, will be 0 everywhere else. Plotting those bounds will result in a parallelogram shape that can be divided into two triangular regions, split across $z = 1$. This creates two cases:

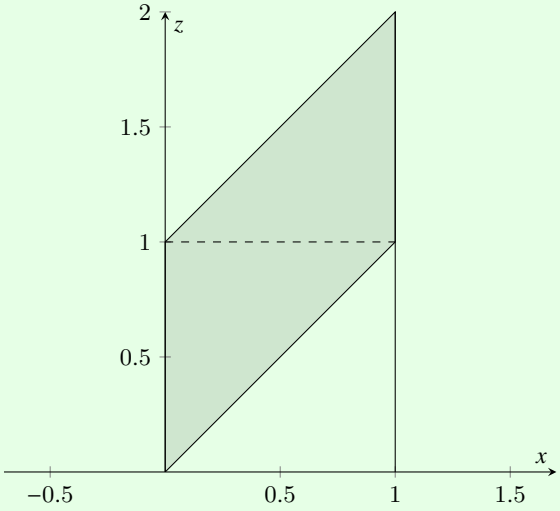
- $0 \le z \le 1$

$$\begin{aligned} f_Z(z) &= \int_0^z 1\,\mathrm{d}x \\ &= z \end{aligned}$$

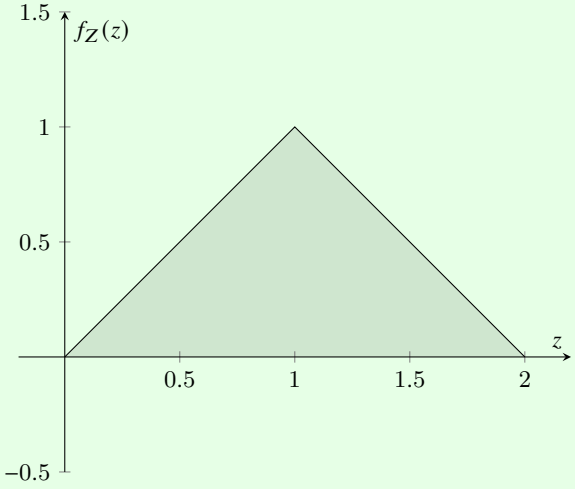
- $1 \le z \le 2$

$$\begin{aligned} f_Z(z) &= \int_{z-1}^1 1\,\mathrm{d}x \\ &= 2 - z \end{aligned}$$

Region bounded by $0 \le x \le 1$ and $0 \le z - x \le 1$



PDF of Z



8 Lecture 8

8.1 Moment Generating Function

Definition 8.1: Moment Generating Function
For a random variable X, its MGF is defined as

$$M_X(t) = \mathbb{E}\left[e^{tX}\right]$$

for $t \in \mathbb{R}$ (provided the expectation exists, which is an assumption we make in this class).

Generating functions transform a sequence $(a_0, a_1, a_2, \dots)$ into the polynomial given by the power series $\sum_{k\geq 0} a_k z^k$. The convolution of such sequences is equivalent to the multiplication of their respective polynomials and multiplying the MGF of two RVs gives the MGF of their sum. Why?

$$\begin{aligned} M_{X+Y}(t) &= \mathbb{E}\left[e^{t(X+Y)}\right] \\ &= \mathbb{E}\left[e^{tX}e^{tY}\right] \end{aligned}$$

Note 8.1
If X and Y are independent, then $f(X)$ and $f(Y)$ are also independent.

Thus,

$$\begin{aligned} M_{X+Y}(t) &= \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right] \\ &= M_X(t)M_Y(t) \end{aligned}$$

Where does the “moment" in moment generating functions come from? Note that

$$\begin{aligned} M_X(t) &= \mathbb{E}\left[e^{tX}\right] \\ &= \mathbb{E}\left[\sum_{n\geq 0}\frac{(tX)^n}{n!}\right] \\ &= \sum_{n\geq 0}\frac{t^n}{n!}\mathbb{E}\left[X^n\right] \end{aligned}$$

In other words, MGFs encode the moments (expectation of the integral powers of an RV) of a distribution into the coefficients of a power series.

8.1.1 Recovering moments from the MGF

If a given MGF exists, it uniquely determines the CDF of X, whose moments can be recovered as follows:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}M_X(t)\bigg|_{t=0} &= \mathbb{E}[X] \\ \frac{\mathrm{d}^2}{\mathrm{d}t^2}M_X(t)\bigg|_{t=0} &= \mathbb{E}\left[X^2\right] \\ &\vdots \\ \frac{\mathrm{d}^n}{\mathrm{d}t^n}M_X(t)\bigg|_{t=0} &= \mathbb{E}\left[X^n\right] \end{aligned}$$

Usually, the distribution of an RV can be extracted by pattern matching its MGF with that of several popular distributions:

- $X \sim \mathcal{N}(\mu, \sigma^2)$ - $M_X(t) = \exp\left(\mu t + \frac{t^2}{2}\sigma^2\right)$ for all $t \in \mathbb{R}$
- $X \sim \text{Exp}(\lambda)$ - $M_X(t) = \frac{\lambda}{\lambda - t}$ for $t < \lambda$
- $X \sim \text{Poisson}(\lambda)$ - $M_X(t) = e^{-\lambda + \lambda e^t}$ for all $t \in \mathbb{R}$
- $X \sim \text{Geometric}(p)$ - $M_X(t) = \frac{pt}{1-(1-p)e^t}$ for $t < -\log(1-p)$
- $X \sim \text{Binomial}(n, p)$ - $M_X(t) = (1 - p + pe^t)^n$
- $X \sim \text{Bernoulli}(p)$ - $M_X(t) = (1-p) + pe^t$

Example 8.1

If $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$. Then,

$$M_X(t) = \exp\left(\mu_x t + \sigma_x^2 \frac{t^2}{2}\right)$$

If X is independent of $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, then

$$\begin{aligned} M_{X+Y}(t) &= \mathbb{E}\left[e^{t(X+Y)}\right] \\ &= \mathbb{E}\left[e^{tX}e^{tY}\right] \\ &= \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right] \\ &= \exp\left(\mu_x t + \sigma_x^2 \frac{t^2}{2}\right)\exp\left(\mu_y t + \sigma_y^2 \frac{t^2}{2}\right) \\ &= \exp\left(t(\mu_x + \mu_y) + \frac{t^2}{2}(\sigma_x^2 + \sigma_y^2)\right) \end{aligned}$$

The expression above is the MGF of $X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

Example 8.2

Let $X_i \sim \text{Bernoulli}(p)$ be IID. Then, $Y = \sum_{i=1}^n X_i \sim \text{Binomial}(n, p)$ and

$$\begin{aligned} M_Y(t) &= \mathbb{E}\left[e^{t(X_1 + \cdots + X_n)}\right] \\ &= \prod_{i=1}^n \mathbb{E}\left[e^{tX_i}\right] \\ &= \prod_{i=1}^n M_{X_i}(t) \\ &= \prod_{i=1}^n (1 - p + pe^t) \\ &= (1 - p + pe^t)^n \end{aligned}$$

8.2 Characteristic Function

In principle, everything that you can do with MGFs can also be done by working directly with distributions. More generally, one usually characterizes functions instead of using MGFs. The characteristic function for an RV X is defined as

$$\phi_X(t) = \mathbb{E}\left[e^{itX}\right]$$

for all $t \in \mathbb{R}$. The term inside the expectation is the Fourier transform of the distribution of X. The characteristic function always exists and just like MGFs, uniquely characterizes a distribution.

8.3 Concentration Inequalities

Seldom in applications can we compute probabilities of interesting events in closed form. Usually, we settle for an inequality. In particular, we usually want to show $\mathbb{P}[A] \approx 0$ or $\mathbb{P}[A] \approx 1$.

8.3.1 Markov's Inequality

Theorem 8.1: Markov Inequality
If X is a non-negative random variable, then

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$$

Proof:

$$\begin{aligned} X &\geq t1_{\{X \geq t\}} \\ \mathbb{E}[X] &\geq t\mathbb{E}\left[1_{\{X \geq t\}}\right] \\ \mathbb{E}[X] &\geq t\mathbb{P}[X \geq t] \\ \mathbb{P}[X \geq t] &\leq \frac{\mathbb{E}[X]}{t} \end{aligned}$$

The first statement can be proven by casework: for $X \geq t$, it is trivially true and for $X \leq t$, the non-negativity of X still makes it true.

Generalized Markov Inequality: if $f(x)$ is a monotonically increasing non-negative function and X is a random variable, then for some $t \geq 0$ and $f(t) > 0$, we have

$$\mathbb{P}[|X| \geq t] = \mathbb{P}[f(|X|) \geq f(t)] \leq \frac{\mathbb{E}[f(|X|)]}{f(t)}$$

8.3.2 Chebyshev's Inequality

Theorem 8.2: Chebyshev's Inequality
If X is a random variable with finite variance, then

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathrm{Var}(X)}{t^2}$$

Proof:

$$\begin{aligned} \mathbb{P}[|X - \mathbb{E}[X]| \geq t] &= \mathbb{P}\left[|X - \mathbb{E}[X]|^2 \geq t^2\right] \\ &\leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|]^2}{t^2} \\ &= \frac{\mathrm{Var}(X)}{t^2} \end{aligned}$$

8.4 Weak Law of Large Numbers

Let $X_i \sim_{\text{IID}} X$. We define the empirical mean as

$$\begin{aligned} M_n &= \frac{X_1 + X_2 + \cdots + X_n}{n} \\ \mathrm{Var}(M_n) &= \mathrm{Var}\left(\frac{X_1 + \cdots + X_2}{n}\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) \end{aligned}$$

$$= \frac{\text{Var}(X)}{n}$$

So, M_n must be "almost constant" for $n >> 1$.

Theorem 8.3: WLLN
Let $X_i \sim X$ and define $M_n = \frac{X_1+\cdots+X_n}{n}$. For $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[|M_n - \mathbb{E}[X]| > \varepsilon] = 0$$
$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| > \varepsilon\right] = 0$$

Proof:

$$\mathbb{E}[M_n] = \mathbb{E}[X] \text{ by LOE}$$
$$\text{Var}(M_n) = \frac{\text{Var}(X)}{n}$$
$$\mathbb{P}[|M_n - \mathbb{E}[X]| > \varepsilon] \leq \frac{\text{Var}(X)}{n\varepsilon^2}$$

Therefore, as $n \to \infty$, then $\frac{\text{Var}(X)}{n\varepsilon^2} \to 0$.

Note 8.2
$\text{Var}(X) < \infty$ is actually not necessary here unlike Chebyshev's Inequality. You just need $\mathbb{E}[X] < \infty$

The empirical frequency of $\{X_i \in B\}$ is given by

$$F_n = \frac{\sum_{i=1}^{n} 1_{\{x_i \in B\}}}{n}$$
$$\mathbb{E}[F_n] = \frac{\sum_{i=1}^{n} P(X_i \in B)}{n}$$
$$= P(X \in B)$$

The weak law of large numbers states that $\mathbb{P}[|F_n - P(X \in B)| > \varepsilon] \to 0$ as $n \to \infty$. In other words, $\mathbb{P}[X \in B]$ is the frequency at which X takes values in B under many repeated trials. Thus, the WLLN justifies probability, in the sense that axiomatic framework is compatible with the "frequentist" or "empirical" framework.

Note 8.3
Our proof of the Weak Law of Large Numbers operated strictly at the level of expectation and variance. We never had to compute the distributions or probabilities explicitly. This is actually fairly typical.

9 Lecture 9

9.1 Chernoff Bounds

Theorem 9.1: Chernoff Bound
For a random variable X, and $a \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}[X \geq a] &\leq \frac{\mathbb{E}\left[e^{tX}\right]}{e^{ta}} \\ &= e^{-ta} M_X(t)\end{aligned}$$

for $t > 0$. Since the MGF is defined for all $t > 0$, we can optimize over $t > 0$ to get a better bound.

Proof:

$$\begin{aligned}\mathbb{P}[X \geq a] &= \mathbb{P}[tX \geq ta] \\ &= \mathbb{P}\left[e^{tX} \geq e^{ta}\right] \\ &\leq \min_t \frac{\mathbb{E}\left[e^{tX}\right]}{e^{ta}}\end{aligned}$$

The main idea behind this: Chebyshev gives a "tighter" bound than Markov because it uses the additional information about the second moment (that $\mathrm{Var}(X) < \infty$). Chernoff, by virtue of the MGF, is using information about all moments of X.

Example 9.1
Let $Z \sim \mathcal{N}(0,1)$. We know that Φ has no closed form expression. But Chernoff gives us good control over the tail probabilities. By Chernoff,

$$\begin{aligned}\mathbb{P}[Z \geq a] &\leq e^{-ta} M_Z(t) \\ &= e^{-ta} e^{\frac{t^2}{2}}\end{aligned}$$

Optimize by setting $t = a$ (using the derivative to minimize $\frac{1}{2}t^2 - ta$) to get

$$\mathbb{P}[Z \geq a] \leq e^{-\frac{a^2}{2}}$$

for $a > 0$.

9.2 Convergence of Random Variables

Convergence of random variables is the language of "limits" in probability. For a sequence of real numbers $a_1, a_2, \cdots \in \mathbb{R}$, we know what it means to write $\lim_{n\to\infty} a_n = a$ (formalized using the epsilon-delta notation covered in introductory calculus/analysis).

Given a sequence of random variables $X_1, X_2, \ldots$, what does it mean to write $\lim_{n\to\infty} X_n = X$? Nothing actually. This is a trick question without any further information. Why? Random variables are functions. For a sequence of functions $f_1, f_2, \ldots$, there are many ways to define the convergence of a sequence of functions $f_1, f_2, \ldots$ (aka $\lim_{n\to\infty} f_n$). Examples include

- Pointwise convergence:

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

  for all x.

- L_1-norm convergence:

  $$\lim_{n\to\infty} \int |f_n(x) - f(x)| \, \mathrm{d}x = 0$$

Since random variables are also functions, we need to specify which type of convergence we are talking about at a particular moment. There are three modes of convergence that we will focus on in this course.

Definition 9.1: Almost Sure Convergence
We say that $X_n \to X$ converges almost surely or with probability one ($X_n \xrightarrow{a.s.} X$ as a shorthand) if

$$\mathbb{P}\left[\lim_{n\to\infty} X_n = X\right] = 1$$
$$\mathbb{P}\left[\left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right] = 1$$

After fixing ω, each $X_n(\omega)$ is just a scalar so their limits are defined. In other words, $X_n \xrightarrow{a.s.} X$ if X pointwise converges to X on some set A of sample points having $P(A) = 1$.

Definition 9.2: Convergence in Probability
We say that $X_n \to X$ converges in probability ($X_n \xrightarrow{p} X$ as a shorthand) if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[|X_n - X| > \varepsilon] = 0$$

Definition 9.3: Convergence in Distribution
We say $X_n \to X$ converges in distribution ($X_n \xrightarrow{d} X$ as a shorthand) if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(n)$$

for all continuity points x of F_X, i.e., where $\mathbb{P}[X = x] = 0$.

Note 9.1
There are notions of convergence in expectation (and different moments) too but that's are not as important for us.

Example 9.2
WLLN tells us that that $\frac{1}{n}\sum_{i=1}^{n} X_n \xrightarrow{p} \mathbb{E}[X]$, i.e the empirical mean converges to the true mean in probability.

Example 9.3
Let $X_i \sim_{\text{IID}} \text{Bernoulli}\left(\frac{1}{2}\right)$. Then, $X_n \to X \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ in distribution (trivially) but X_n does not converge to anything almost surely. Why? In order to converge to something, we need a sequence with finitely many ones or zeroes, which occurs with zero probability. Say you have a sequence of coin flips. Then almost sure convergence would imply that after a finite sequence of coin flips, you're stuck at some value infinitely, which defies logic.

Example 9.4
For a continuous RV X, if $X_n \to X$ pointwise on all points except for $X(\omega) = x$, then $X_n \xrightarrow{a.s.} X$ because $\mathbb{P}[\Omega \setminus \{X = x\}] = \mathbb{P}[\Omega] - \mathbb{P}[X = x] = 1 - 0 = 1$.

9.2.1 Hierarchy of Convergence

Theorem 9.2
If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{p} X$.

Proof: Fix some $\varepsilon > 0$. Define $A_n = \bigcup_{m=n}^{\infty} \{|X_m - X| > \varepsilon\}$ to be the event that X_n deviates from X by more than ε in the future. Observe that $A_{i+1} \subseteq A_i$ for all i, which indicates that these events are "decreasing" (i.e. $\mathbb{P}[A_{i+1}] \leq \mathbb{P}[A_i]$) to some event A_∞ with $\mathbb{P}[A_\infty] = 0$. This follows from the definition of almost sure convergence which states that $X_n(\omega)$ will deviate from $X(\omega)$, for all sample points ω, by an amount more than ε only finitely many times. Thus, as $n \to \infty$, the sequence of real

numbers $|X_n(\omega) - X(\omega)|$ is eventually bounded by ε. Therefore,

$$\lim_{n\to\infty} \mathbb{P}[|X_n(\omega) - X(\omega)| > \varepsilon] \leq \lim_{n\to\infty} \mathbb{P}[A_n] \xrightarrow{a.s.} 0$$

The inequality above follows from the fact that $|X_n(\omega) - X(\omega)| < \varepsilon$ captures only one deviation but A_n captures all possible deviations past n.

The main distinction between almost sure convergence and convergence in probability is the frequency of deviations. For convergence in probability, the probability of deviations will decrease, but it will still be non-zero, allowing for it to occur infinitely many times. On the other hand, almost sure convergence guarantees that only finitely many deviations occur, after which none shall be observed any longer.

Theorem 9.3
If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$.

Proof: Let x be a point such that $\mathbb{P}[X = x] = 0$. Fix some $\varepsilon > 0$. Then, the law of total probability states that

$$\begin{aligned}
\mathbb{P}[X_n \leq x] &= \mathbb{P}[X_n \leq x \cap X \leq x + \varepsilon] + \mathbb{P}[X_n \leq x \cap X > x + \varepsilon] \\
&\leq \mathbb{P}[X \leq x + \varepsilon] + \mathbb{P}[X_n - X \leq x - X \cap x - X < -\varepsilon] \\
&\leq \mathbb{P}[X \leq x + \varepsilon] + \mathbb{P}[X_n - X < -\varepsilon] \\
&\leq \mathbb{P}[X \leq x + \varepsilon] + \mathbb{P}[X_n - X < -\varepsilon] + \mathbb{P}[X_n - X > \varepsilon] \\
&= \mathbb{P}[X \leq x + \varepsilon] + \mathbb{P}[|X_n - X| > \varepsilon]
\end{aligned}$$

Similarly,

$$\mathbb{P}[X \leq x - \varepsilon] \leq \mathbb{P}[X_n \leq x - \varepsilon + \varepsilon] + \mathbb{P}[|X - X_n| > \varepsilon] = \mathbb{P}[X_n \leq x] + \mathbb{P}[|X_n - X| > \varepsilon]$$

Combining the bounds above,

$$\mathbb{P}[X \leq x - \varepsilon] - \underbrace{\mathbb{P}[|X_n - X| \geq \varepsilon]}_{\to 0} \leq \mathbb{P}[X_n \leq x] \leq \mathbb{P}[X \leq x + \varepsilon] + \underbrace{\mathbb{P}[|X_n - X| \geq \varepsilon]}_{\to 0}$$
$$\mathbb{P}[X \leq x - \varepsilon] \leq \mathbb{P}[X_n \leq x] \leq \mathbb{P}[X \leq x + \varepsilon]$$

The above follows from the definition of convergence in probability: $\mathbb{P}[|X_n - X| \leq \varepsilon] = 1 \implies \mathbb{P}[|X_n - X| > \varepsilon] = 0$. Since $\mathbb{P}[X = x] = 0$ as the CDF of X is continuous at x, the bounds above hold for all $\varepsilon > 0$ as $n \to \infty$. Therefore, letting $\varepsilon \to 0$ will give us $\mathbb{P}[X_n \leq x] \to \mathbb{P}[X \leq x]$.

In summary, the hierarchy of convergence looks like the following:

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$$

These implications are strict, i.e., almost sure convergence will always imply convergence in probability which will always imply convergence in distribution, but the other directions are not necessarily true.

9.3 Strong Law of Large Numbers

Theorem 9.4: SLLN
If $X_1, X_2, \ldots \sim_{\text{IID}} X$, and $\mathbb{E}[X] < \infty$, then,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mathbb{E}[X]$$

It is the same statement as the WLLN, but claims almost sure convergence instead of just convergence in probability. Since it is a stronger statement, the WLLN can essentially be ignored entirely now as it is nothing more than a corollary of the strong law. However, the SLLN is also considerably harder to prove.

The SLLN tells us that for a sample point ω, the empirical mean $\frac{1}{n}\sum_{i=1}^{n} X_i(\omega)$ converges to $\mathbb{E}[X]$ for all ω in some event having probability 1 (which does not have to be the entire Ω). Essentially, if you observe all the sample paths starting at ω, i.e. the sequences $(X_1(\omega), X_2(\omega), \cdots)$ for IID X_i, then all of the probability will start getting concentrated around those with mean $\mathbb{E}[X]$ and all other sequences will become "anomalies" that occur with probability 0. The SLLN does not tell us that anomalies can't happen, but it tells us that they won't happen.

It is possible to have a sequence of coin flips where all flips land on heads. So this anomaly can happen. However, the SLLN tells us that we will never observe such a sequence with probability 1.

9.4 Central Limit Theorem

Theorem 9.5: Central Limit Theorem
Let $X_1, \ldots, X_n \sim_{\text{IID}} X$ such that $\text{Var}(X) = \sigma^2 < \infty$ and $\mathbb{E}[X] = \mu$. Define $S_n = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n}\sigma}$. Then, $S_n \xrightarrow{d} Z \sim \mathcal{N}(0,1)$. In other words, $\lim_{n\to\infty} \mathbb{P}[X_n \le x] = \Phi(x), \forall x \in \mathbb{R}$.

Proof: WLOG, let $\text{Var}(X) = 1$ and $\mathbb{E}[X] = 0$.

$$\begin{aligned} M_X(t) &= \sum_{n\ge 0} t^n \frac{\mathbb{E}[X^n]}{n!} \\ &= 1 + \frac{t^2}{2} + o(t^2) \end{aligned}$$

The little o notation above indicates that $o(t^2)$ is a series of terms that is vanishing faster than t^2 as $t^2 \to \infty$. Since t is fixed, the expression $o(t^2)$ can be rewritten as $o\left(\frac{1}{n}\right)$ since the n term is actually what will make the higher order terms disappear below:

$$\begin{aligned} \lim_{n\to\infty} M_{S_n}(t) &= \lim_{n\to\infty} M_{\frac{X_1+\cdots+X_n}{\sqrt{n}}}(t) \\ &= \lim_{n\to\infty} \left(M_{\frac{X}{\sqrt{n}}}(t)\right)^n \\ &= \lim_{n\to\infty} \left(M_X\left(\frac{t}{\sqrt{n}}\right)\right)^n \\ &\approx \lim_{n\to\infty} \left(1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \\ &\approx e^{\frac{1}{2}t^2} \end{aligned}$$

since $\lim_{n\to\infty}\left(1 + \frac{c}{n}\right)^n = e^c$. Thus, $\lim_{n\to\infty} M_{S_n}(t) = M_{\mathcal{N}(0,1)}(t)$, which implies that $\lim_{n\to\infty} S_n \xrightarrow{d} \mathcal{N}(0,1)$.

Note 9.2
Pointwise convergence of characteristic functions implies convergence in distribution is a non-trivial fact known as Levy's Continuity Theorem.

Note 9.3
$X_n \to X$ almost surely does not imply $\mathbb{E}[X_n] \to \mathbb{E}[X]$. We need more information to conclude this.

10 Lecture 10

10.1 Information Theory

The origin of information theory can be traced back to a paper by Claude Shannon in 1948 called the "Mathematical Theory of Communication." This singular paper kicked off the "information age" and addressed some of the following questions

- How reliably and how quickly can I communicate a message over a noisy channel? Ex. the cocktail party problem. This is also called the channel coding problem.
- How many bits do I need to losslessly represent an observation? Ex. data compression. This is also called the source coding problem.

Shannon's mathematical insights guided decades of development into these areas (that are still ongoing).

10.2 Source Coding (i.e. Compression)

Definition 10.1: Entropy

For a discrete random variable $X \sim P_X$, we define its (Shannon) entropy as

$$\begin{aligned} H(X) &= \sum_x P_X(x) \log\left(\frac{1}{P_X(x)}\right) \\ &= \mathbb{E}\left[\log \frac{1}{P_X}\right] \\ &= -\mathbb{E}[\log P_X(X)] \end{aligned}$$

where the log is taken in base 2 (entropy is measured in units of "bits").

The function $f(x) = -\log P_X(x)$ describes the "uncertainty" associated with observing $X = x$. This is in agreement with our intuition since $f(x)$ is larger when $P_X(x)$ is a smaller (we are more surprised when a low probability value occurs) and vice versa. Thus, the entropy $H(X)$ is the expected uncertainty of X, i.e., "how random" X is on average (just like $\mathrm{Var}(X)$ is another metric that describes the "randomness" in X via its spread).

The interpretation of $H(X)$ as "uncertainty" is further justified by the source coding theorem, which says that $H(X)$ is the number of bits needed to describe X on average (equivalent to the number of yes/no questions to determine X, on average).

Theorem 10.1: Source Coding Theorem

For any $\varepsilon > 0$, discrete RVs $X_i \sim_{\text{IID}} P_X$ can be losslessly represented using $\leq n(H(X) + \varepsilon)$ bits. Conversely, any representation, using $< nH(X)$ bits is impossible without loss of information.

The result has two parts:

1. descriptions $\leq n(H(X) + \varepsilon)$ bits are possible
2. descriptions $< nH(X)$ bits are not possible

Therefore, the entropy of an RV X is a fundamental asymptotic limit on its compression. So, in a way, it describes the information content of X.

Example 10.1: Huffman Codes

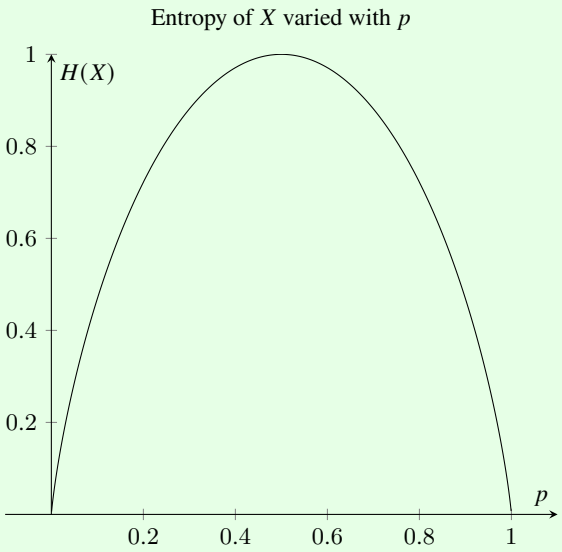
They take in a sequence $X_1, X_2, \ldots, X_n \sim_{\text{IID}} P_X$, and output a string of bits $\approx nH(X)$ in length (on average).

Example 10.2

For $X \sim \text{Bernoulli}(p)$, let X be a coin flip. The entropy of X is $H(X) = -p \log p - (1-p)\log(1-p)$. Now, consider the

following cases:

- If I flip a fair ($p = \frac{1}{2}$) coin n times, I need n bits to represent all outcomes.
- If I flip a biased ($p \approx 0.11$) coin n times, then I only need $\approx \frac{n}{2}$ bits to describe all outcomes.
- If I flip a biased ($p = 0$) coin n times, I need 0 bits to describe all outcomes.



How is the second case possible? Concentration: from a lot of randomness comes determinism.

Example 10.3: Entropy vs Variance

Let $X \sim \text{Bernoulli}(p)$ and $Y = aX$. Then,

$$\begin{aligned}\text{Var}(X) &= p(1-p)\\ \text{Var}(Y) &= a^2\,\text{Var}(X)\\ &= a^2 p(1-p)\end{aligned}$$

In the case above, the variance of X and Y differ even though they are both distributed with the same bias, i.e., we are equally "surprised" by the values of X and Y. Therefore, for this purpose, entropy is a better measure of the randomness in X and Y since it is only considers their probability distributions but not their numerical values.

10.3 Properties of Entropy

Since $P_X(x) \in [0,1] \implies \frac{1}{P_X(x)} \geq 1 \implies \log \frac{1}{P_X(x)} \geq 0 \implies H(X) = \mathbb{E}\left[\log \frac{1}{P_X(x)}\right] \geq 0.$

Theorem 10.2: Jensen's Inequality

For any random variable X and convex function $f(\cdot)$, Jensen's inequality states that $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$. Flip the direction of the inequality if $f(\cdot)$ is concave.

Let $\mathcal{X}$ be the set of values that X can take. Since $\log(x)$ is concave,

$$\begin{aligned}H(X) &= \mathbb{E}\left[\log \frac{1}{P_X}\right]\\ &\leq \log \mathbb{E}\left[\frac{1}{P_X}\right]\end{aligned}$$

$$= \log\left(\sum_{x\in\mathcal{X}} \frac{1}{P_X(x)} P_X(x)\right)$$
$$= \log|\mathcal{X}|$$

Jensen's inequality is applicable here since $\frac{1}{P_X(X)}$ is also a random variable (since it is the function $P_X(\cdot)^{-1}$ applied to X).

Definition 10.2: Joint Entropy
The joint entropy of two RVs X and Y is

$$H(X,Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_{XY}(x,y)}$$
$$= \mathbb{E}\left[\log \frac{1}{P_{XY}(X,Y)}\right]$$

The joint entropy of X and Y can be interpreted as the information content of both X and Y.

Definition 10.3: Conditional Entropy
The conditional entropy of X given Y is defined as

$$H(X \mid Y) = \mathbb{E}\left[\log \frac{1}{P_{X|Y}(X \mid Y)}\right]$$
$$= \sum_y P_Y(y) H(X \mid Y = y)$$
$$= \sum_y P_Y(y) \mathbb{E}\left[\log \frac{1}{P_{X|Y}(X \mid Y = y)}\right]$$
$$= \sum_y P_Y(y) \sum_x P_{X|Y}(x \mid y) \log \frac{1}{P_{X|Y}(x \mid y)}$$

This quantity can be interpreted as the information content provided by X given Y, i.e., the information provided by X that is not already given by Y.

When X and Y are independent (i.e. $P_{X|Y}(x \mid y) = P_X(x)$), summation above reduces to

$$H(X \mid Y) = \sum_y P_Y(y) \sum_x P_X(x) \log \frac{1}{P_X(x)}$$
$$= \left(\sum_y P_Y(y)\right) H(X)$$
$$= H(X)$$

Theorem 10.3: Chain Rule of Entropy
$H(X,Y) = H(X \mid Y) + H(Y)$

Proof: Following the definition of conditional entropy,

$$H(X \mid Y) = \sum_y P_Y(y) \sum_x P_{X|Y}(x \mid y) \log \frac{1}{P_{X|Y}(x \mid y)}$$
$$= \sum_{x,y} P_Y(y) P_{X|Y}(x \mid y) \log \frac{P_Y(y)}{P_Y(y) P_{X|Y}(x \mid y)}$$
$$= \sum_{x,y} P_{XY}(x,y) \log \frac{P_Y(y)}{P_{XY}(x,y)}$$

$$
\begin{aligned}
&= \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_{XY}(x,y)} - \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_Y(y)} \\
&= \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_{XY}(x,y)} - \sum_{y} P_Y(y) \log \frac{1}{P_Y(y)} \sum_x P_{Y|X}(y \mid x) \\
&= \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_{XY}(x,y)} - \sum_{y} P_Y(y) \log \frac{1}{P_Y(y)} \\
&= H(X,Y) - H(Y)
\end{aligned}
$$

Thus, $H(X,Y) = H(X \mid Y) + H(Y)$. The chain rule is essentially saying that the information content of both X and Y is the sum of the information provided by Y and the information provided by X that is not already given by Y.

10.4 Asymptotic Equipartition Theorem

For a sequence $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ where all $X_i \sim_{\text{IID}} P_X$, let the probability of observing it be

$$\mathbb{P}[X_1, X_2, \ldots, X_n] = \prod_{i=1}^{n} P_X(x_i)$$

For convenience, define the notation that the probability of the sequence $(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)$ occurring is

$$P_X(x_1)P_X(x_2)\ldots P_X(x_n) = P_X(X_1)P_X(X_2)\ldots P_X(X_n)$$

Theorem 10.4: Aymptotic Equipartition Theorem

If $(X_i)_{i \geq 1} \sim_{\text{IID}} P_X$, then,

$$-\frac{1}{n} \log \mathbb{P}[X_1, X_2, \ldots, X_n] \xrightarrow{p} H(X)$$

In other words, with overwhelming probability

$$\mathbb{P}[X_1, X_2, \ldots, X_n] \approx 2^{-nH(X)}$$

Proof: Using WLLN,

$$
\begin{aligned}
-\frac{1}{n} \log \mathbb{P}[X_1, \ldots, X_n] &= \frac{1}{n} \log \prod_{i=1}^{n} \frac{1}{P_X(X_i)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \log \frac{1}{P_X(X_i)} \\
&\xrightarrow{p} \mathbb{E}\left[\frac{1}{P_X(X)}\right] \\
&= H(X)
\end{aligned}
$$

11 Lecture 11

11.1 Applying AEP to Source Coding

Definition 11.1: Typical Set
Fix $\varepsilon > 0$, and for each $n \geq 1$, define the "typical set"

$$A_\varepsilon^{(n)} = \left\{(X_1, \ldots, X_n) : \mathbb{P}[X_1, \ldots, X_n] \geq 2^{-n(H(X)+\varepsilon)}\right\}$$

This set is a subset of all the possible observable sequences.

Some properties of the typical set include

1. $\mathbb{P}\left[(X_1, \ldots, X_n) \in A_\varepsilon^{(n)}\right] \to 1$ as $n \to \infty$ by AEP (more precisely, by WLLN).

   Proof:

   $$\begin{aligned}\mathbb{P}\left[(X_1, \ldots, X_n) \notin A_\varepsilon^{(n)}\right] &= \mathbb{P}\left[\mathbb{P}[X_1, \ldots, X_n] < 2^{-n(H(X)+\varepsilon)}\right] \\ &= \mathbb{P}\left[-\frac{1}{n}\log \mathbb{P}[X_1, \ldots, X_n] > H(X) + \varepsilon\right] \\ &= \mathbb{P}\left[\left|-\frac{1}{n}\log \mathbb{P}[X_1, \ldots, X_n] - H(X)\right| > \varepsilon\right] \\ &\xrightarrow{p} 0\end{aligned}$$

   by the AEP (WLLN).

2. $\left|A_\varepsilon^{(n)}\right| \leq 2^{n(H(X)+\varepsilon)}$

   Proof:

   $$\begin{aligned}1 &\geq \sum_{(X_1,\ldots,X_n)\in A_\varepsilon^{(n)}} \mathbb{P}[X_1, \ldots, X_n] \\ &\geq \sum_{(X_1,\ldots,X_n)\in A_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)} \\ &= \left|A_\varepsilon^{(n)}\right| 2^{-n(H(X)+\varepsilon)}\end{aligned}$$

   The first line follows from the definition of the typical set as a subset.

The two properties are essentially saying that while the typical set is exponentially smaller in size with respect to the set of all possible sequences (property 2), for large n, virtually all of the probability mass is concentrated around it (property 1). This is simply just another way of rephrasing the Asymptotic Equipartition Property described before.

We will now use the AEP to prove the "achievability" part of the source coding theorem. How many bits, at most, do I need to represent N objects? Should be $\log N$ at most. Then, form the following protocol for source coding:

- If I observe $(X_1, \ldots, X_n) \in A_{\frac{\varepsilon}{2}}^{(n)}$, I will describe it using $\approx \log\left|A_{\frac{\varepsilon}{2}}^{(n)}\right| \leq n\left(H(X) + \frac{\varepsilon}{2}\right)$ bits by property 2
- If I observe $(X_1, \ldots, X_n) \notin A_{\frac{\varepsilon}{2}}^{(n)}$, I just describe it with brute force using $n \log |X|$ bits.

For this protocol, what is the average description length?

$$\mathbb{E}[\text{no. of bits}] \leq n\left(H(X) + \frac{\varepsilon}{2}\right) \underbrace{\mathbb{P}\left[(X_1, \ldots, X_n) \in A_{\frac{\varepsilon}{2}}^{(n)}\right]}_{\leq 1 \text{ by property 1}} + \underbrace{n \log |X| \cdot \mathbb{P}\left[(X_1, \ldots, X_n) \notin A_{\frac{\varepsilon}{2}}^{(n)}\right]}_{\leq \frac{\varepsilon}{2}}$$

$$\leq n(H(X)+\varepsilon) \text{ for all } n \text{ sufficiently large}$$

The second expression is vanishing as n increases, so it can be bounded by any positive constant: $\frac{\varepsilon}{2}$ is the most convenient one since it makes the algebra work out. This concludes the proof of achievability. The converse proof is omitted.

11.2 Information Transmission (Channel Coding)

How do we send information reliably over an unreliable channel? Suppose we fix some "rate" $R > 0$. Let there be a message $M \sim \text{Uniform}(\{1, \dots, 2^{nR}\})$, which takes nR bits to represent since $H(M) = nR$.

$$M \xrightarrow{\text{encoder}} X^n(M) \xrightarrow{\text{noisy channel}} Y^n \xrightarrow{\text{decoder}} \hat{M}(Y^n)$$

Outline of the process above:

- The message M goes through some encoder, which outputs the vector $X^n(M) = (X_1(M), \dots, X_n(M))$.
- This sequence goes through a noisy channel and gets corrupted to $Y^n(M) = (Y_1, \dots, Y_n)$. Intuitively, each Y_i is a "noisy version" of X_i
- The decoder takes the noisy sequence and recovers the message $\hat{M}(Y_n)$.

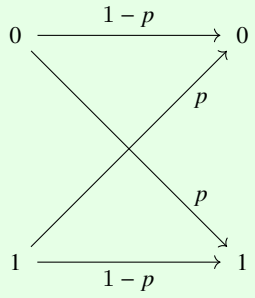
The parameters associated with this system include

- The rate $R = \frac{H(M)}{n} = \frac{\text{\# of "information bits"}}{\text{\# of channel uses}}$
- The error probability $P_e^{(n)} = \max \mathbb{P}\left[\hat{M} \neq M\right]$

In other words, R can be viewed as the measure of a channel's bandwidth and P_e as the measure of reliability.

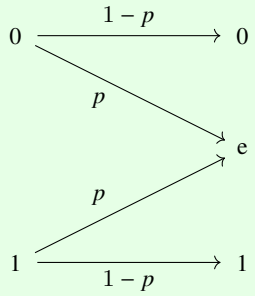
Example 11.1: Binary Symmetric Channel

A BSC(p) flips a bit independently with probability p:



Example 11.2: Binary Erasure Channel

A BEC(p) erases a bit independently with probability p:



Note 11.1
Channels, in general, are represented by a conditional PMF $P_{Y|X}$.

Example 11.3
BSC(p) is represented by

$$P_{Y|X}(y \mid x) = \begin{cases} p & y \neq x \\ 1-p & y = x \end{cases}$$

Definition 11.2: Mutual Information
For a channel $P_{Y|X}$ and input distribution P_X, there is also the

$$\begin{aligned} P_{XY} &= P_X(x)P_{Y|X}(y \mid x) && \text{(joint distribution of inputs and outputs)} \\ P_Y &= \sum_x P_{XY}(x,y) && \text{(marginal distribution of outputs)} \end{aligned}$$

The mutual information $I(X;Y)$ is defined as

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}$$

This measure is a function of the channel distribution $P_{Y|X}$ and the input distribution P_X only. It can be interpreted as the information gained about X (the input) from observing Y (the output).

Note 11.2
Mutual information is symmetric, i.e., $I(X;Y) = I(Y;X)$. Observe that

$$\begin{aligned} I(X;Y) &= \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \\ &= \sum_{x,y} P_{XY}(x,y) \left(\log \frac{1}{P_X(x)} - \log \frac{P_Y(y)}{P_{XY}(x,y)} \right) \\ &= \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_X(x)} - \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_{X|Y}(x \mid y)} \\ &= H(X) - H(X \mid Y) \end{aligned}$$

Moreover, substituting $H(X \mid Y) = H(X,Y) - H(Y)$ gives $I(X;Y) = H(X) + H(Y) - H(X,Y)$. Note how $I(Y;X)$ is equivalent to the same quantity.

Definition 11.3: Channel Capacity
For a channel $P_{Y|X}$, its capacity is defined as

$$C = \max_{P_X} I(X;Y)$$

In other words, it is the maximum mutual information between the channel inputs and outputs over all possible input distributions.

Like mutual information, the channel capacity is also a function of just the channel distribution $P_{Y|X}$. It is also the maximum rate at which information can be reliably transmitted without a high probability of error. Consequently, the goal of current information theory research is to develop channels with higher capacities and transmit information with greater rates: this will increase the mutual information gained about the inputs from the outputs, which is what reliable communication ultimately is all about.

Note 11.3
The capacity has no dependence on the number of bits transmitted!

Theorem 11.1: Shannon's Channel Coding Theorem
Fix channel $P_{Y|X}$ and $\varepsilon > 0$ and $R < C$.

- For all n sufficiently large, there exists a rate R communication scheme (encoder/decoder) that achieves $P_e^{(n)} < \varepsilon$.
- If $R > C$, then $P_e^{(n)} \to 1$ for any sequence of communication schemes.

Example 11.4
The capacity of a BSC(p) is $C = 1 - H_2(p)$ where $H_2(p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$

Example 11.5
The capacity of a BEC(p) is $C = 1 - p$

We will now prove the Channel Coding theorem for the special case of BEC(p). We need to show two things:

- Any rate $R > C$ is not possible
- All rates $R < C$ allow for reliable communication

Proof: We will prove each part separately.

- Part 1

  Consider a block of n channel uses. The transmitter does not know what locations will be erased. However, let's suppose that a genie told us. Say $M = (0, 1, 1, \ldots, 0)$ is the message to be sent. Then, knowing the positions of the bits that will be erased, the transmitter can send the following:

  $$0 \quad 1 \quad e \quad e \quad 1 \quad e \quad \ldots \quad 0 \quad e$$

  The transmitter can send information without error in un-erased positions.

  How many un-erased positions are there? Less than $n(1 - p + \varepsilon)$ with overwhelming probability for any $\varepsilon > 0$ and all n sufficiently large. Thus, the transmitter can only reliably send $\approx n(1-p)$ bits, which implies that $R \leq (1-p)$.

- Part 2

  This relies on the "probabilistic method": don't actually construct an explicit scheme and analyze it, but just be lazy and show that one exists with a non-zero probability. Fix some $\varepsilon > 0$ and $R < 1 - p - \varepsilon$. Then, generate a random matrix

  $$\mathbf{C} = \begin{bmatrix} C_{1,1} & C_{1,2} & \ldots & C_{1,n} \\ C_{2,1} & C_{2,2} & \ldots & C_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{2^{nR},1} & C_{2^{nR},2} & \ldots & C_{2^{nR},1} \end{bmatrix}$$

  for $C_{ij} \sim \text{Bernoulli}\left(\frac{1}{2}\right)$. Give $\mathbf{C}$ to both the encoder and the decoder. Define the following protocol:

  1. On observing the message $M \in \left\{1, \ldots, 2^{nR}\right\}$, send the Mth row of $\mathbf{C}$, i.e., $X^n(M) = (C_{M,1}, \ldots, C_{M,n}) \in \{0,1\}^n$
  2. On receiving Y^n, look for row in $\mathbf{C}$ that matches (modulo erasures). In other words, find a row that matches all the un-erased packets (the index of this row will be the intended message).

  We can only error if ≥ 2 rows match what was received.

12 Lecture 12

12.1 Information Transmission (Channel Coding) Cont.

Example 12.1
Example of the protocol described last time. Let $n = 4$ and $R = \frac{1}{2}$. Then, let

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$$M = 3$$
$$X^n(M) = (0, 1, 1, 0)$$
$$Y^n = (0, 1, e, e)$$

Then, the decoder will locate $\hat{M} = 3$ as the only row that starts with $(0, 1)$. An error could occur if more than one row in $\mathbf{C}$ is consistent with Y^n. Consider $Y^n = (0, 0, e, e)$ in the example above. This could correspond to either $M = 1$ or $M = 2$.

Continuing the proof of the setup from last time, show that probability of error averaged over choices of $\mathbf{C}$ is small for large n.

Proof: Let $E \subset \{1, 2, \ldots, n\} = [n]$ be the positions that are erased by the channel. WLOG (by symmetry), assume that $M = 1$ is correct. Then,

$$\begin{aligned} \mathbb{E}\left[P_e^{(n)}\right] &= \mathbb{E}\left[1_{\{\hat{M} \neq M\}}\right] \\ &= \sum_{E \subset [n]} \mathbb{E}\left[1_{\{\hat{M} \neq M\}} \mid E\right] \mathbb{P}[\text{bits erased} = E] \\ &\leq \left(\sum_{E : |E| \leq n\left(p + \frac{\varepsilon}{2}\right)} \mathbb{E}\left[1_{\{\hat{M} \neq M\}} \mid E\right] \mathbb{P}[E] \right) + \mathbb{P}\left[|E| > n\left(p + \frac{\varepsilon}{2}\right)\right] \end{aligned}$$

The upper bound splits the expression based on the number of erasures (the size of E). Note that due to the law of large numbers $\mathbb{P}\left[|E| > n\left(p + \frac{\varepsilon}{2}\right)\right] = \mathbb{P}\left[\frac{|E|}{n} - p > \frac{\varepsilon}{2}\right] \xrightarrow{p} 0$. Thus, the expression on the right will be very small and can essentially be ignored. So, for $|E| \leq n\left(p + \frac{\varepsilon}{2}\right)$, look at

$$\mathbb{E}\left[1_{\{\hat{M} \neq M\}} \mid E\right] = \mathbb{E}\left[1_{\{\hat{M} \neq 1\}} \mid E\right] = \mathbb{P}\left[\bigcup_{m \geq 2}^{2^{nR}} \{C(1, [n] \setminus E) = C(m, [n] \setminus E)\} \mid E\right]$$

The expression $C(i, [n] \setminus E)$ denotes row i of $\mathbf{C}$, with the columns indexed by E removed. The equality of these rows will imply that some row m matches row 1 as well, which will lead to an error. The rows will only match if the $n - |E|$ positions that were not erased are identical (which are each independent with $p = \frac{1}{2}$). Using the union bound,

$$\begin{aligned} \mathbb{E}\left[1_{\{\hat{M} \neq 1\}} \mid E\right] &\leq \sum_{m \geq 2}^{2^{nR}} \left(\frac{1}{2}\right)^{n - |E|} \\ &\leq 2^{nR - (n - |E|)} \\ &\leq 2^{-n\frac{\varepsilon}{2}} \end{aligned}$$

since $nR - n + |E| \leq n(1 - p - \varepsilon) - n + n\left(p + \frac{\varepsilon}{2}\right) = -n\frac{\varepsilon}{2}$.
Substituting this bound,

$$\mathbb{E}\left[P_e^{(n)}\right] \leq \left(\sum_{E : |E| \leq n\left(p + \frac{\varepsilon}{2}\right)} 2^{-n\frac{\varepsilon}{2}} \mathbb{P}[E] \right) + \mathbb{P}\left[|E| > n\left(p + \frac{\varepsilon}{2}\right)\right]$$

$$\leq 2^{-n\frac{\varepsilon}{2}} + \mathbb{P}\left[\frac{|E|}{n} > \left(p + \frac{\varepsilon}{2}\right)\right]$$
$$\to 0$$

as $n \to \infty$. Thus, there must exist some choice of codebook and n sufficiently large to make $P_e^{(n)} < \varepsilon$.

12.2 Markov Chains

Random variables by themselves are only so interesting. Often, we are interested in sequences of random variables $(X_n)_{n\geq 0}$ (called random or stochastic processes). They can be used to model real life things such as

- Robot position over time
- Website visited by internet use
- Signal received by cell tower

Up until now, the only processes we have seen have been IID which is a good starting point for nice results (such as WLLN, SLLN, CLT) but these are limited for modeling real scenarios. "Markov chains" are one level above IID processes. They are a very flexible class of processes and useful for modeling a wide variety of situations.

Definition 12.1: Markov Chain
$(X_n)_{n\geq 0}$ is a Markov chain if each X_i is a discrete RV taking values in a discrete set S (state space) and for all $n \geq 0$ and $i, j \in S$,

$$\mathbb{P}[X_{n+1} = j \mid X_n = i, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0] = \mathbb{P}[X_{n+1} = j \mid X_n = i]$$

i.e. the future only depends on the past through the present - every single future state is only dependent on the current state and nothing else! This is also called the Markov property.

Any process with finite memory will satisfy the Markov property above. The state space can be augmented as S^w (cartesian product of S) instead of just S to be represented as w-tuples of states for w units (steps) of memory.

Definition 12.2: State
X_n is called the "state" of the process at time $n \geq 0$.

Note 12.1
If $(X_n)_{n\geq 0}$ is a MC, then there is an implied underlying probability space $(\Omega, \mathcal{F}, P)$ on which all X_n's are random variables. This just follows from the Kolmogorov Extension Theorem mentioned way earlier.

Definition 12.3: Temporally Homogeneous Markov Chains
$\mathbb{P}[X_{n+1} = j \mid X_n = i] = P_{ij}$ for all $i, j \in S$ and $n \geq 0$. The "transition probabilities" P_{ij} from i to j are not dependent on time.

Transition probabilities must satisfy a few rules:

1. $P_{ij} \geq 0, \forall i, j \in S$
2. $\sum_{j\in S} P_{ij} = 1, \forall i \in S$

There is a direction associated with transition probabilities to P_{ij} is not the same as P_{ji}. The transition probabilities $(P_{ij})_{i,j\in S}$ describe the statistics of an MC.

Definition 12.4: Transition Matrix

It is helpful to store the transition probabilities in a matrix form as follows

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} & \dots \\ P_{21} & P_{22} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

where $[\mathbf{P}]_{ij} = P_{ij}$. We say the transition matrix is stochastic, i.e., all entries are non-negative and rows sum to 1.

Markov chains can be represented by "state transition diagrams" where

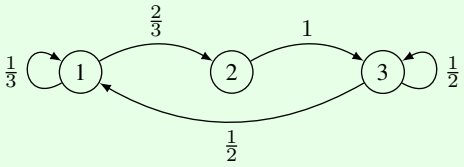
- Each state represented by a node
- Arrows between states represent transitions and are labeled with the transition probabilities

Example 12.2

Let

$$\mathbf{P} = \begin{bmatrix} 1/3 & 2/3 & 0 \\ 0 & 0 & 1 \\ 1/2 & 0 & 1/2 \end{bmatrix}$$

Its state transition diagram is given by



The transition matrix $\mathbf{P}$ lists the "one-step" transition probabilities. What about $\mathbb{P}[X_n = j \mid X_0 = i]$, which is an "n-step transition probability"? Denote it as P_{ij}^n (notation! this is an n-step transition probability and not P_{ij} raised to the nth-power).

Theorem 12.1: Chapman-Kolmogorov Equations

The n step transition probabilities can be computed as

$$P_{ij}^n = [\mathbf{P}^n]_{ij}$$

where $\mathbf{P}^n$ is the transition matrix $\mathbf{P}$ raised to the nth power.

Proof: Induct on n:

- Base case $n = 1$ holds by previous definitions
- Induction Step

$$\begin{aligned} \mathbb{P}[X_{n+1} = j \mid X_0 = i] &= \sum_{k \in S} \mathbb{P}[X_{n+1} = j, X_n = k \mid X_0 = i] \\ &= \sum_{k \in S} \mathbb{P}[X_{n+1} = j \mid X_n = k, X_0 = i] \mathbb{P}[X_n = k \mid X_0 = i] \\ &= \sum_{k \in S} \mathbb{P}[X_{n+1} = j \mid X_n = k] \mathbb{P}[X_n = k \mid X_0 = i] \\ &= [\mathbf{P} \times \mathbf{P}^n]_{ij} \\ &= [\mathbf{P}^{n+1}]_{ij} \end{aligned}$$

We applied the Markov property to convert $\mathbb{P}[X_{n+1} = j \mid X_n = k, X_0 = i]$ to $\mathbb{P}[X_{n+1} = j \mid X_n = k]$, which is the jth column of $\mathbf{P}$. Moreover, by the inductive hypothesis, $\mathbb{P}[X_n = k \mid X_0 = i]$ gives the ith row of $\mathbf{P}^n$. Their dot product, i.e. the sum above, gives the ijth element of $\mathbf{P}^{n+1}$.

13 Lecture 13

13.1 Classification of States

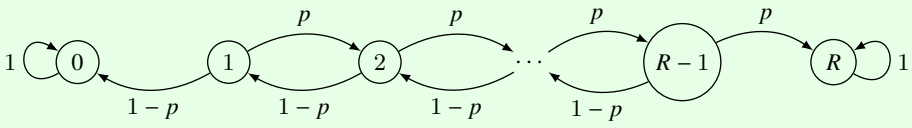
If there is a path in the state-transition diagram from i to j (i.e. $P_{ij}^n > 0$ for some $n \geq 1$), then we say that j is accessible from i and we write $i \to j$. If we also have $j \to i$, then we write $i \leftrightarrow j$ (states i, j "communicate"). By convention, $i \leftrightarrow i, \forall i \in S$. We claim that $\leftrightarrow$ is an equivalence relation on S since

- Reflexive: $i \leftrightarrow i, \forall i \in S$
- Symmetric: $i \leftrightarrow j \iff j \leftrightarrow i, \forall i, j \in S$
- Transitive: $i \leftrightarrow k, k \leftrightarrow j \implies u \leftrightarrow j, \forall i, j, k \in S$

Therefore, the equivalence relation $\leftrightarrow$ partitions S into "equivalence classes" of communicating states. If $C \subseteq S$ is a "class" and $i \in C$, then $j \in C \iff i \leftrightarrow j$.

Example 13.1: Gambler's Ruin

For $p \in (0, 1)$, consider



The three classes of states in this example are $\{0\}, \{1, 2, \ldots, R-1\}, \{R\}$.

13.2 Class Properties

Definition 13.1: Irreducible

A MC is irreducible if it has only one class (i.e. the entire state space S).

Definition 13.2: Recurrent

A state $i \in S$ is said to be recurrent if, given that $X_0 = i$, the process revisits i with probability one.

Note 13.1

Recurrence is the same as saying I will visit state i infinitely many times with probability one, given that I reach state i at least once.

Definition 13.3: Transient

A state $i \in S$ is transient if it is not recurrent.

Note 13.2

Transience is the same as saying I will visit state i only finitely many times, given that I start in state i, after which I may never return back to it in the future. A simple example is that of a state that has at least one transition out of its own class to some other class.

Example 13.2

In the gambler's ruin chain above, states $\{0\}$ and $\{R\}$ are recurrent while $\{1, \ldots, R-1\}$ is transient.

Note 13.3
Recurrence and Transience are class properties. (i.e. if C is a class and $i \in C$ is transient, then all $j \in C$ are transient. The same holds for recurrence.)

Proof: It suffices to show that if i recurrent, then so is j. In particular, it suffices to show that if $X_0 = i$, then I will land in j after finite time with probability 1. Since $i \leftrightarrow j$, $\exists n \geq 1$ such that $P_{ij}^n > 0$. So, I will land in j after $X \sim \text{Geometric}(P_{ij}^n)$ visits to i. This is equivalent to tossing a coin with probability P_{ij}^n. If the coin turns up 0, then we don't land in j. However, since i is being visited infinitely often, I keep getting more and more tries, that are independent of each other by the Markov property. Finally, flipping a coin can be modeled using geometric random variables.

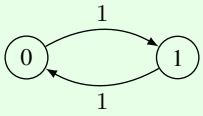
Definition 13.4: Positive and Null Recurrent
Define the random variable $T_i = \min\{n \geq 1 : X_n = i\}$, which is the first time $n \geq 1$ that state i is entered in. If $i \in S$ is recurrent, we further classify it as

- positive recurrent if $\mathbb{E}[T_i \mid X_0 = i] < \infty$
- null recurrent if $\mathbb{E}[T_i \mid X_0 = i] = \infty$

Note 13.4
Positive and null recurrence are also class properties. The proof should follow similarly from the one above.

Definition 13.5: Period
For $i \in S$, let $\text{period}(i) = \text{GCD}\{n \geq 1 : P_{ii}^n > 0\}$. In other words, if I start in state i, then revisits to state i only occur at integer multiples of $\text{period}(i)$.

Example 13.3
Consider the following MC:

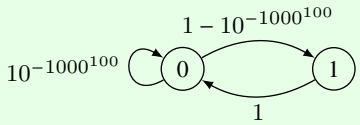


Observe that period(0) = period(1) = 2.

Note 13.5
Periodicity is a class property, i.e., if $i \leftrightarrow j$, then $\text{period}(i) = \text{period}(j)$.

Definition 13.6: Aperiodic
An irreducible MC is aperiodic if any state (and therefore all states) has period 1.

Example 13.4
Aperiodicity is a "brittle" property. Even adding a self loop with a small probability, such as in



will make the MC above aperiodic.

Definition 13.7: Stationary Distribution

A probability distribution $\boldsymbol{\pi} = (\pi(i))_{i \in S}$, which can be represented by a row vector, is said to be a stationary (or invariant) distribution iff

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$$
$$\pi_j = \sum_{i \in S} \pi_i P_{ij}$$

for all $j \in S$. These are called the balance equations.

Why is it called a stationary distribution? If $X_0 \sim \pi$, then $X_n \sim \pi$ for all $n \geq 0$. In other words, distribution over states is invariant in time, and the resulting process $(X_n)_{n \geq 0}$ is stationary (not changing over time).

13.3 Long Term Behavior

Theorem 13.1: Big MC Theorem

Let $(X_n)_{n \geq 0}$ be an irreducible MC. Exactly one of the following true:

1. Either all states are transient or all are null recurrent. In this case, no stationary distribution exists and
$$\lim_{n \to \infty} P_{ij}^n = 0, \forall i, j \in S$$
2. All states are positive recurrent. In this case, a stationary distribution $\boldsymbol{\pi}$ exists. It is unique and satisfies
$$\pi_j = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} P_{ij}^k = \frac{1}{\mathbb{E}\left[T_j \mid X_0 = j\right]}$$
Moreover, if the MC is aperiodic,
$$\lim_{n \to \infty} P_{ij}^n = \pi_j, \forall i, j \in S$$

Note 13.6

Every irreducible finite state MC is positive recurrent, i.e., only the second option of the Big Theorem is possible for it.

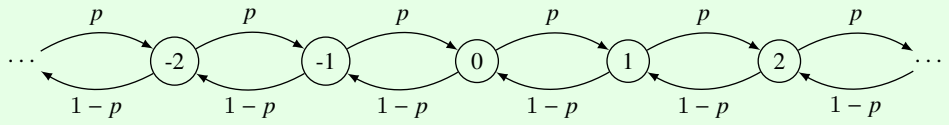
Note 13.7

The second part of the Big Theorem is in fact a biconditional: an irreducible Markov chain is positive recurrent if and only if a unique stationary distribution exists.

What if the MC is not irreducible? Then it will have different communicating classes within it, not all of which will be positive recurrent (think about it). For those of you who have taken CS 170, you can essentially view this as a metagraph of the Markov chain, where each SCC is a different class. If there are multiple positive recurrent classes, they will each behave like their own separate irreducible positive recurrent MC with their own stationary distributions. Thus, the SD of the overall chain will be some convex combination of these separate individual SDs, and hence, will not be unique.

Example 13.5: Random Walk on Integers

Consider a random walk on $\mathbb{Z}$ with transition probability $p \in (0, 1)$, represented by

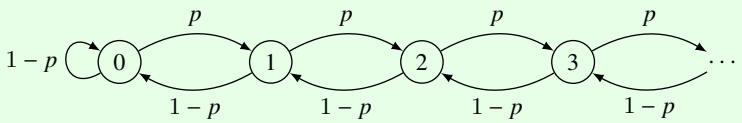


This chain is irreducible (all of the states are communicating). Moreover,

- If $p = \frac{1}{2}$, then all states are null recurrent (using CLT).
- If $p \neq \frac{1}{2}$, then all states are transient (using SLLN).

Example 13.6: Birth-Death Chain

Consider the following Markov chain:



This chain is

- positive recurrent if $p < \frac{1}{2}$
- null recurrent if $p = \frac{1}{2}$
- transient if $p > \frac{1}{2}$

Example 13.7: Page Rank

Model the internet by a finite directed graph $G = (V, E)$. Let $(X_n)_{n \geq 0}$ be a random walk on this graph, where next state chosen uniformly from outgoing links. We can use Markov chains to model this and rank the important of webpages: by the Big Theorem, a stationary distribution exists for this MC, and π_j describes the fraction of traffic on site j in the long run. So, π_j is a good proxy for the "rank" of page j.

14 Lecture 14

14.1 Recap of Lecture 13

Example 14.1
Let $(X_n)_{n\geq 0}$ be a random walk on the hypercube $\{0,1\}^n$, where the next vertex is chosen by randomly flipping one of the current state's bits. If we start at state $(0,\dots,0)$, what is the expected number of steps before returning to it? By the Big Theorem,

$$\mathbb{E}\left[T_{(0,\dots,0)} \mid X_0 = (0,\dots,0)\right] = \frac{1}{\pi_{(0,\dots,0)}} = \frac{1}{2^{-n}} = 2^n$$

where $\pi_{(0,\dots,0)} = 2^{-n}$ follows from symmetry.

Example 14.2
Suppose we collect reward $R(i)$ on entering state $i \in S$. If R is bounded and the MC is irreducible and positive recurrent, then,

$$\frac{1}{n}\sum_{k=1}^{n} R(X_k) \to \mathbb{E}[R(X)]$$

almost surely for $X \sim \pi$. The term on the LHS is the average reward collected over the first n states. Why? Let $N_j(n)$ be the number of entries into j up to time n. Then,

$$\frac{1}{n}\sum_{k=1}^{n} R(X_k) = \sum_{j\in S} \frac{N_j(n)}{n} R(j) \xrightarrow{SLLN} \sum_{j\in S} \frac{1}{\mathbb{E}\left[T_j \mid X_0 = j\right]} R(j) = \sum_{j\in S} \pi_j R_j$$

14.2 Reversibility

The Big Theorem describes when a stationary distribution can exist but how do we compute it in the first place? Solving $\pi_j = \sum_i \pi_i P_{ij}$ for all j is one way. It turns out this is much easier when the MC is "reversible."

Definition 14.1: Reversible
An irreducible MC is reversible if there exists a probability vector $\boldsymbol{\pi}$ satisfying

$$\pi_j P_{ji} = \pi_i P_{ij}$$

for all $i, j \in S$. These are called the detailed balanced equations (DBEs).

Note 14.1
If an MC is reversible, then $\boldsymbol{\pi}$ is a stationary distribution (in fact, unique by irreducibility from the Big Theorem) since

$$\pi_j P_{ji} = \pi_i P_{ij} \implies \pi_j \sum_i P_{ji} = \sum_i \pi_i P_{ij} \implies \pi_j = \sum_i \pi_i P_{ij} \implies \boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$$

So, in the case of a reversible MC, just solve the detailed balanced equations to find π.

Where does the term "reversible" come from? If we start a reversible MC where $X_0 \sim \boldsymbol{\pi}$, then

$$(X_0, X_1, \dots, X_n) \overset{d}{=} (X_n, X_{n-1}, \dots, X_0)$$

Reversible Markov chains show up quite frequently in practice because physical systems that have a good notion of equilibrium are typically reversible.

Proof: Here is a quick proof of the backwards Markov property: start at some state $X_{k+n} = x_{k+n}$ and move "forwards" in the reverse direction (i.e., backwards in time) to state $X_k = x_k$. Then,

$$\mathbb{P}[X_k = x_k \mid X_{k+1} = x_{k+1}, \dots, X_{k+n} = x_{k+n}]$$

$$= \frac{\mathbb{P}[X_k = x_k, X_{k+1} = x_{k+1}, \ldots, X_{k+n} = x_{k+n}]}{\mathbb{P}[X_{k+1} = x_{k+1}, \ldots, X_{k+n} = x_{k+n}]}$$
$$= \frac{\mathbb{P}[X_k = x_k]\mathbb{P}[X_{k+1} = x_{k+1} \mid X_k = x_k]\mathbb{P}[X_{k+2} = x_{k+2} \mid X_{k+1} = x_{k+1}] \ldots \mathbb{P}[X_{k+n} = x_{k+n} \mid X_{k+n-1} = x_{k+n-1}]}{\mathbb{P}[X_{k+1} = x_{k+1}]\mathbb{P}[X_{k+2} = x_{k+2} \mid X_{k+1} = x_{k+1}] \ldots \mathbb{P}[X_{k+n} = x_{k+n} \mid X_{k+n-1} = x_{k+n-1}]}$$
$$= \frac{\mathbb{P}[X_k = x_k]\mathbb{P}[X_{k+1} = x_{k+1} \mid X_k = x_k]}{\mathbb{P}[X_{k+1} = x_{k+1}]}$$

The transition from the second to third line follows from the chain rule of conditional probability, along with the regular Markov property. Assuming we start at the stationary distribution here,

$$\frac{\mathbb{P}[X_k = x_k]\mathbb{P}[X_{k+1} = x_{k+1} \mid X_k = x_k]}{\mathbb{P}[X_{k+1} = x_{k+1}]} = \frac{\pi_{x_k} P_{x_k, x_{k+1}}}{\pi_{x_{k+1}}}$$

Notice that the probability of being at the next state ($X_k = x_k$) only depends on itself and the current state ($X_{k+1} = x_{k+1}$), making the reversed chain a valid MC as well!

Thus, if a given MC is reversible, i.e., the original chain behaves like its reversed version, then P_{ij} will be the same transition probability in both chains. In other words,

$$P_{ji} = \frac{\pi_i P_{ij}}{\pi_j} \implies \pi_j P_{ji} = \pi_i P_{ij}$$

as desired.

Note 14.2
The detailed balanced equations are sufficient but not necessary for proving the existence of an SD. If a distribution satisfies the DBEs, then it will be an SD but the converse is not always true.

Example 14.3
If a Markov chain is a tree (imagine starting with an undirected graph that is a tree, but replacing each undirected edge (u, v) with directed edges (u, v) and (v, u), and adding any self-loops as needed), then it will satisfy the detailed balance equations.

Example 14.4
Consider a random walk on an undirected graph $G = (V, E)$, i.e., V is the state space, E is the set of transitions, and the next state from a given state is chosen uniformly at random among its neighbors. Then, the stationary distribution for this MC is given by

$$\pi_i = \frac{\deg(i)}{\sum_{j \in V} \deg(j)} = \frac{\deg(i)}{2|E|}$$

This follows because it satisfies the DBEs:

$$\pi_i P_{ij} = \frac{\deg(i)}{2|E|} \cdot \frac{1}{\deg(i)} = \frac{1}{2|E|} = \frac{\deg(j)}{2|E|} \cdot \frac{1}{\deg(j)} = \pi_j P_{ji}$$

14.3 First Step Analysis

The Big Theorem tells us about the asymptotic behavior of an irreducible MC. We will now turn our attention to techniques for analyzing the finite-horizon behavior of (not necessarily irreducible) MCs.

Definition 14.2: Hitting Times
Consider $A \subset S$, and define the hitting time $T_A = \min\{n \geq 0 : X_n \in A\}$. Therefore, T_A is a random variable

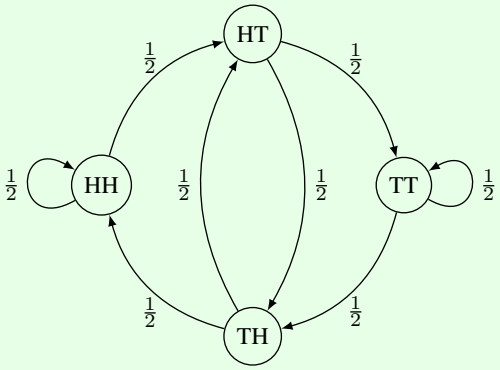
The distribution of T_A is very hard to compute so let's take a look at its expectation instead. The general strategy for calculating expected hitting time is

- Define
$$t_i = \mathbb{E}[T_A \mid X_0 = i]$$
This is the expected amount of time to hit A, when starting from state i.
- Formulate the "first-step" equations recursively:
  - For $i \notin A$,
$$\mathbb{E}[T_A \mid X_0 = i] = 1 + \sum_{j \in S} P_{ij}\mathbb{E}[T_A \mid X_0 = j]$$
The sum is a consequence of the law of total expectation - after transitioning to state j with probability P_{ij}, you are starting in that state from scratch again. The additional 1 accounts for the actual step taken from state i to state j.
  - For $i \in A$, let $\mathbb{E}[T_A \mid X_0 = i] = 0$ because i is already contained in A.

Summarizing,
$$\text{FSE} = \begin{cases} t_i = 1 + \sum_j P_{ij} t_j & i \notin A \\ t_i = 0 & i \in A \end{cases}$$

Example 14.5

Consider tossing a fair coin. How many tosses will it take until we get 2 tails in a row?



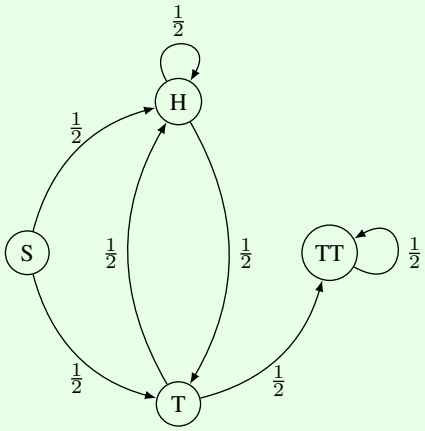
Given above is the MC to track the two most recent tosses. Letting states 1 be HH, 2 be HT, 3 be TH and 4 be TT, the transition matrix is
$$\mathbf{P} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \end{bmatrix}$$
Start in state HH, i.e., compute $t_{\text{HH}} = \mathbb{E}[T_{\text{TT}} \mid X_0 = \text{HH}]$. Set up the FSEs to get
$$\begin{aligned} t_{\text{HH}} &= 1 + \frac{1}{2}t_{\text{HH}} + \frac{1}{2}t_{\text{HT}} \\ t_{\text{HT}} &= 1 + \frac{1}{2}t_{\text{TH}} + \frac{1}{2}t_{\text{TT}} \\ t_{\text{TH}} &= 1 + \frac{1}{2}t_{\text{HT}} + \frac{1}{2}t_{\text{HH}} \\ t_{\text{TT}} &= 0 \end{aligned}$$
On solving, $t_{\text{HH}} = 6$, $t_{\text{HT}} = 4$ and $t_{\text{TH}} = 6$.

15 Lecture 15

15.1 First Step Analysis Cont.

Example 15.1

Flip a fair coin until two tails occur, at which point stop. What is the expected number of heads that I see?



Setting up the modified FSEs,

$$\begin{aligned} t_i &= \mathbb{E}[\text{no. heads seen} \mid X_0 = i], \forall i \in S \\ t_S &= \frac{1}{2}t_H + \frac{1}{2}t_T \\ t_H &= 1 + \frac{1}{2}t_H + \frac{1}{2}t_T \\ t_T &= \frac{1}{2}t_H + \frac{1}{2}t_{TT} \\ t_{TT} &= 0 \end{aligned}$$

Solving this system will yield $t_S = 3$, $t_H = 4$ and $t_T = 2$.

Consider states $A, B \subset S$ where $A \cap B = \emptyset$. Let,

$$\begin{aligned} T_A &= \min\{n \geq 0 : X_n \in A\} \\ T_B &= \min\{n \geq 0 : X_n \in B\} \end{aligned}$$

Definition 15.1: Hitting Probabilities

The hitting probability is defined as $\mathbb{P}[T_A < T_B \mid X_0 = i]$.

Define

$$\alpha(i) = \mathbb{P}[T_A < T_B \mid X_0 = i]$$

Then, our FSEs will be

$$\text{FSE} = \begin{cases} \alpha(i) = 0 & i \in B \\ \alpha(i) = 1 & i \in A \\ \alpha(i) = \sum_{j \in S} P_{ij}\alpha(j) & i \notin A \cup B \end{cases}$$

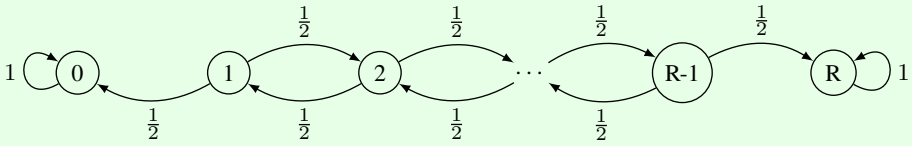
The last equation comes from

$$\mathbb{P}[T_A < T_B \mid X_0 = i] = \sum_j P_{ij}\mathbb{P}[T_A < T_B \mid X_0 = j]$$

which is basically the law of total probability.

Example 15.2

Consider the following Gambler's Ruin chain with an equal chance of losing or gaining money (same transition probability between preceding and succeeding states):



We claim that $\mathbb{P}[T_R < T_0 \mid X_0 = i] = \frac{i}{R}$. Why? Assume that $\alpha(i)$ is indeed $\frac{i}{R}$. Then, setting up the FSEs will also yield

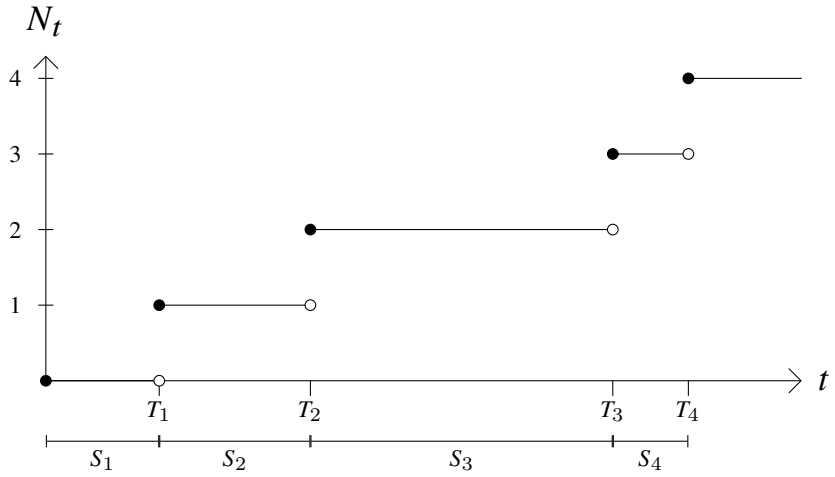
$$\begin{aligned}
\alpha(0) &= 0 \\
\alpha(R) &= 1 \\
\alpha(i) &= \frac{1}{2}\frac{(i-1)}{R} + \frac{1}{2}\frac{(i+1)}{R} \qquad i \in \{1, \dots, R-1\} \\
&= \frac{i}{R}
\end{aligned}$$

15.2 Poisson Processes

Definition 15.2: Counting Process

A counting process $(N_t)_{t \geq 0}$ is a continuous-time integer-valued random process, which has right-continuous sample paths (similar to how CDFs were defined).

A Poisson Process is an example of a counting process, which forms the basis for continuous time Markov chains. Graphically, this process can look like the following:



The times $T_i = \min\{t \geq 0 : N_t \geq i\}$ are called the "arrival times" of each event that is counted. The gap between two consecutive arrival times is called the "inter-arrival time", denoted by $S_i = T_i - T_{i-1}$ for $i \geq 1$ (and $T_0 = 0$).

Definition 15.3: Poisson Process

A rate λ Poisson Process, denoted by $\mathrm{PP}(\lambda)$, is a counting process with inter-arrival times $S_i \sim_{\mathrm{IID}} \mathrm{Exp}(\lambda)$.

Why "Poisson"?

Theorem 15.1
If $(N_t)_{t\geq 0}$ is a $\mathrm{PP}(\lambda)$, then for each $t \geq 0$, $N_t \sim \mathrm{Poisson}(\lambda t)$. In other words, $\mathbb{P}[N_t = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$

Proof:

$$\begin{aligned}
\mathbb{P}[N_t = n] &= \mathbb{P}[T_n \leq t < T_{n+1}] \\
&= \mathbb{E}\left[1_{\{T_n \leq t\}} 1_{\{t < T_n + S_{n+1}\}}\right] \\
&= \int f_{T_n}(s) 1_{\{s \leq t\}} \mathbb{E}\left[1_{\{t < s + S_{n+1}\}}\right] \mathrm{d}s \\
&= \int_0^t f_{T_n}(s) \mathbb{E}\left[1_{t-s < S_{n+1}}\right] \mathrm{d}s \\
&= \int_0^t f_{T_n}(s) \mathbb{P}[S_{n+1} \geq t - s] \, \mathrm{d}s \\
&= \int_0^t f_{T_n}(s) e^{-\lambda(t-s)} \, \mathrm{d}s
\end{aligned}$$

The sum of n IID exponentials is the "Erlang" distribution. Substituting its PDF in the expression above,

$$\begin{aligned}
\mathbb{P}[N_t = n] &= \int_0^t \lambda \frac{e^{-\lambda s}(\lambda s)^{n-1}}{(n-1)!} e^{-\lambda(t-s)} \, \mathrm{d}s \\
&= \frac{\lambda^n}{(n-1)!} e^{-\lambda t} \int_0^t s^{n-1} \, \mathrm{d}s \\
&= \frac{\lambda^n}{(n-1)!} e^{-\lambda t} \frac{t^n}{n} \\
&= \frac{(\lambda t)^n e^{-\lambda t}}{n!}
\end{aligned}$$

Observe that by the memoryless property of $\mathrm{Exp}(\lambda)$, if $(N_t)_{t\geq 0} \sim \mathrm{PP}(\lambda)$, then $(N_{t+s} - N_s)_{t\geq 0}$ is also a $\mathrm{PP}(\lambda)$ for all $s \geq 0$. Moreover, $(N_{t+s} - N_s)_{t\geq 0}$ is independent of $(N_\tau)_{0<\tau<s}$. In particular, Poisson processes have a Markov Property (the future is independent of the past).

Poisson process also have independent and stationary increments, i.e., if $t_0 < t_1 < t_2 < \cdots < t_k$, then the increments $(N_{t_1} - N_{t_0}), (N_{t_2} - N_{t_1}), \ldots, (N_{t_k} - N_{t_{k-1}})$ are independent and $N_{t_i} - N_{t_{i-1}} \sim \mathrm{Poisson}(\lambda(t_i - t_{i-1}))$ for all i. Stationary refers to the increments being chosen independently of previous times - shifting the entire process in time will not change its behavior.

15.3 Conditional Distribution of Arrivals

Theorem 15.2
Conditioned on $\{N_t = n\}$,

$$(T_1, T_2, \ldots, T_n) \overset{d}{=} (U^{(1)}, U^{(2)}, \ldots, U^{(n)})$$

where each $U^{(i)}$ is the order statistic of n Uniform(0, t) random variables. In other words, given n arrivals occurred up to time t, their arrival times look like IID Uniform(0, t) random variables in distribution.

Proof: For $0 = t_0 \leq t_1 \leq \cdots \leq t_n \leq t$, we have using Bayes Rule

$$\begin{aligned}
f_{T_1, T_2, \ldots T_n | N_t}(t_1, \ldots, t_n \mid N_t = n) &= \frac{\mathbb{P}[N_t = n \mid T_1 = t, \ldots, T_n = t_n]}{\mathbb{P}[N_t = n]} f_{T_1, \ldots, T_n}(t_1, \ldots, t_n) \\
&= \frac{\mathbb{P}\left[N_t - N_{t_n} = 0 \mid T_1 = t_1, \ldots, T_n = t_n\right]}{\mathbb{P}[N_t = n]} \prod_{i=1}^n f_{S_i}(t_i - t_{i-1}) \\
&= \frac{\mathbb{P}\left[N_t - N_{t_n} = 0\right]}{\mathbb{P}[N_t = n]} \prod_{i=1}^n f_{S_i}(t_i - t_{i-1})
\end{aligned}$$

$$= \frac{e^{-\lambda(t-t_n)}}{e^{-\lambda t}\frac{(\lambda t)^n}{n!}} \prod_{i=1}^{n} \lambda e^{-\lambda(t_i - t_{i-1})}$$
$$= \frac{n!}{t^n}$$

The conditioning on $T_1 = t_1, \ldots T_n = t_n$ goes away because these increments are independent and stationary. Moreover, the Uniform distribution on $(0,t)^n$ has density $\frac{1}{t^n}$ on $[0,t]^n$. Order statistics sort among $n!$ permutations, so we get $n!$ multiples on the region where coordinates are sorted.

Example 15.3

Suppose cars pass through a tollbooth and follow a $\mathrm{PP}(\lambda)$ where λ has units of vehicles/minute.

1. What is the probability that no vehicles pass in 2 minutes?

$$\mathbb{P}[N_2 = 0] = e^{-2\lambda}$$

2. What is the expected number of vehicles to pass in 2 minutes?

$$\mathbb{E}[N_2] = 2\lambda$$

3. Given 10 vehicles passed in 2 minutes, what is the expected number that passed in the first 30 seconds?

$$\frac{10}{4} \quad \left(\mathbb{E}\left[\sum_{i=1}^{10} 1_{u_i \le \frac{1}{2}}\right] \text{ for } u_i \sim \mathrm{Uniform}(0,2)\right)$$

Example 15.4

Suppose photons arrive at a detector $\sim \mathrm{PP}(\lambda)$. If 10^6 photons are detected in 2 seconds, what is the distribution of the number of photons that arrived in the first second? Conditioned on the number of arrivals, each arrival time will look like its uniformly distributed over $(0,2)$. Therefore, the probability of arrival in the interval $(0,1)$ is 0.5. Since all arrivals are independent, the number of photons in the first second $\sim \mathrm{Binomial}(10^6, 0.5)$.

16 Lecture 16

16.1 Recap

A counting process $(N_t)_{t\geq 0}$ is a $\text{PP}(\lambda)$ iff

1. $N_0 = 0$
2. $N_t - N_s \sim \text{Poisson}(\lambda(t-s))$ for $0 \leq s \leq t$
3. $(N_t)_{t\geq 0}$ has independent increments

16.2 Merging and Splitting/Thinning

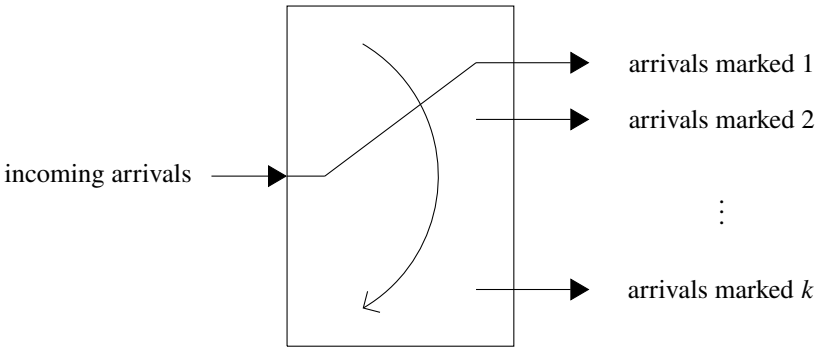
Theorem 16.1: Merging
If $(N_{1,t}) \sim \text{PP}(\lambda_1)$, $(N_{2,t}) \sim \text{PP}(\lambda_2)$ are independent, then $(N_{1,t} + N_{2,t}) \sim \text{PP}(\lambda_1 + \lambda_2)$.

Proof: We will verify the three conditions given above:

1. $N_{1,0} + N_{2,0} = 0 + 0 = 0$
2. $(N_{1,t} + N_{2,t}) - (N_{1,s} + N_{2,s}) = (N_{1,t} - N_{1,s}) + (N_{2,t} - N_{2,s}) \overset{d}{=} \text{Poisson}(\lambda_1(t-s)) * \text{Poisson}(\lambda_2(t-s))$. However, this convolution will return a poisson RV with the two rates added, i.e., $\text{Poisson}((\lambda_1 + \lambda_2)(t-s))$
3. $(N_{1,t} + N_{2,t})_{t\geq 0}$ should have independent increments but this is true since $(N_{1,t})_{t\geq 0}$ and $(N_{2,t})_{t\geq 0}$ also have independent increments

Example 16.1
Suppose men are hospitalized by COVID according to $\sim \text{PP}(\lambda_1)$, independent of women who are hospitalized $\sim \text{PP}(\lambda_2)$. Then, the total number of patients hospitalized $\sim \text{PP}(\lambda_1 + \lambda_2)$.

Let $p_1, p_2, \ldots, p_k$ be probabilities such that $\sum_{i=1}^{k} p_i = 1$. Let $(N_t)_{t\geq 0}$ be a $\text{PP}(\lambda)$. Mark each arrival of this process with a label that is sampled according to these probabilities (i.e., label i with probability p_i), independently of all the other arrivals. Let $(N_{i,t})_{t\geq 0}$ be the process that counts the arrivals with label i for $i = 1, \ldots, k$.



Random switch that is in position i with probability p_i

Theorem 16.2: Splitting/Thinning
$(N_{i,t})_{t\geq 0}$ for $i = 1, \ldots, k$ are independent Poisson processes with respective rates of $p_i\lambda$.

Proof: We can consider just $k = 2$ (since any more events are basically a result of splitting apart further). Let $p_1 = p$ and $p_2 = 1 - p$.

$$\begin{aligned}
\mathbb{P}\left[N_{1,t} = n, N_{2,t} = m\right] &= \mathbb{P}\left[N_{1,t} = n, N_{2,t} = m, N_t = n + m\right] \\
&= \mathbb{P}\left[N_{1,t} = n, N_{2,t} = m \mid N_t = n + m\right]\mathbb{P}[N_t = n + m] \\
&= \binom{n+m}{n} p^n (1-p)^m e^{-\lambda t} \frac{(\lambda t)^{n+m}}{(n+m)!} \\
&= \underbrace{e^{-p\lambda t} \frac{(p\lambda t)^n}{n!}}_{\text{PMF for Poisson}(p\lambda t}) \times \underbrace{e^{-(1-p)\lambda t} \frac{((1-p)\lambda t)^m}{m!}}_{\text{PMF for Poisson}((1-p)\lambda t)} \\
&= \mathbb{P}\left[N_{1,t} = n\right]\mathbb{P}\left[N_{2,t} = m\right]
\end{aligned}$$

Thus, $N_{1,t} \sim \text{Poisson}(p\lambda t)$ and $N_{2,t} \sim \text{Poisson}((1-p)\lambda t)$ are independent.

Example 16.2
Suppose packets arrive to a router $\sim \text{PP}(\lambda)$. They are randomly routed to outgoing link A with probability p, and outgoing link B with probability $(1-p)$. Then,

$$\begin{aligned}
\text{packets on link A} &\sim \text{PP}(p\lambda) \\
\text{packets on link B} &\sim \text{PP}((1-p)\lambda)
\end{aligned}$$

and both processes are independent.

Note 16.1
These properties seem simple. However, they can be extremely powerful for problem solving.

Example 16.3
Let $A \sim \text{PP}(\lambda)$ and $B \sim \text{PP}(\mu)$. Merging A and B and splitting the result with probabilities $\frac{\lambda}{\lambda+\mu}$ and $\frac{\mu}{\lambda+\mu}$ respectively will return the original processes. This tells us that if A and B are going on simultaneously, then the probability that an arrival occurs at A before B is $\frac{\lambda}{\lambda+\mu}$.

16.3 Random Incidence Paradox

Consider $(N_t)_{t\geq 0} \sim \text{PP}(\lambda)$. Suppose I pick a "random" time t_0. What is the expected length of the inter-arrival interval in which t_0 falls? Conventional wisdom would say that it is $\frac{1}{\lambda}$. Say t_0 falls between arrivals T_i and T_{i+1}. Furthermore,

$$L = (T_{i+1} - t_0) + (t_0 - T_i)$$

Note that $T_{i+1} - t_0 \sim \text{Exp}(\lambda)$ by the memoryless property of the exponential distribution. Then,

$$\begin{aligned}
\mathbb{P}[t_0 - T_i > s] &= \mathbb{P}[\text{No arrivals in interval } (t_0 - s, t_0)] \\
&= \mathbb{P}\left[N_{t_0 - s} - N_{t_0} = 0\right] \\
&= e^{\lambda(t_0 - s - t_0)} \\
&= e^{-\lambda s}
\end{aligned}$$

Thus, $t_0 - T_i \sim \text{Exp}(\lambda)$ as well. By the linearity of expectation, $\mathbb{E}[L] = \frac{1}{\lambda} + \frac{1}{\lambda} = \frac{2}{\lambda}$, which is twice the average inter-arrival time! What is the explanation? If we arrive at a random time, we are more likely to end up landing in a long interval over a short one.

16.4 Continuous Time Markov Chains

Consider representing a $\mathrm{PP}(\lambda)$ as follows:

We start in state 0, wait for an $\mathrm{Exp}(\lambda)$ amount of time before transitioning. Then repeat. If $(X_t)_{t\geq 0}$ is a process where X_t is the state at time $t \geq 0$, then $(X_t)_{t\geq 0}$ is a $\mathrm{PP}(\lambda)$. This is an example of a CTMC. In fact, all CTMCs look sort of like this.

Similar to DTMCs, we assume a countable state space S for CTMCs. Recall that DTMCs are defined by a transition matrix $\mathbf{P}$ — since there is a time component in CTMCs, we will define those using a rate matrix instead.

Definition 16.1: Rate Matrix

A CTMC is defined in terms of a rate matrix $\mathbf{Q}$ satisfying

1. $[\mathbf{Q}]_{ij} \geq 0$ for $i \neq j$ and $i, j \in S$
2. $\sum_{j\in S}[\mathbf{Q}]_{ij} = 0$ for all $i \in S$

i.e. off diagonal elements of $\mathbf{Q}$ are non-negative and each row of $\mathbf{Q}$ sums to 0.

Note 16.2

$[\mathbf{Q}]_{ii} = -\sum_{j\neq i}[\mathbf{Q}]_{ij}$

Definition 16.2: Transition Rate

For convenience, we define $q_i = -[\mathbf{Q}]_{ii} = \sum_{j\neq i}[\mathbf{Q}]_{ij}$ as the "transition" rate for state i.

Note that one can rewrite

$$[\mathbf{Q}]_{ij} = q_i P_{ij}$$

for all i, j for some $(P_{ij})_{i,j\in S}$ that satisfies

$$\sum_{j\in S} P_{ij} = 1$$

where $P_{ii} = 0$ and $P_{ij} \geq 0$.

Definition 16.3: Jump Chain

These P_{ij}'s are transition probabilities for an associated DTMC called the "jump chain".

The way a CTMC with rate matrix $\mathbf{Q}$ works is as follows:

- Start with $X_0 = i$
- Hold for an $\mathrm{Exp}(q_i)$ amount of time, then jump to state $j \in S$ with probability P_{ij}
- Hold for an $\mathrm{Exp}(q_j)$ amount of time, then jump to state $k \in S$ with probability P_{jk}
- Repeat

If X_t is the state at time $t \geq 0$, then $(X_t)_{t\geq 0}$ is a CTMC. Why is it a Markov chain? By the memoryless property of an exponential distribution, $\mathbb{P}[X_{t+\tau} = j \mid X_t = i, X_s = i_s, 0 \leq s < t] = \mathbb{P}[X_{t+\tau} = j \mid X_t = i]$.

Note 16.3

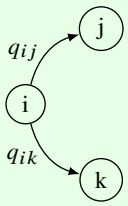
Any continuous time process with the above Markov property can be realized using the above procedure.

Observe that a CTMC and, thus, its associated jump (or embedded) chain does not contain any self-loops. In DTMCs, with discrete time intervals, a self-loop would indicate that the MC left and re-entered the same state. However, in CTMCs, with

continuous time, those intervals are not well-defined anymore so the concept of leaving and re-entering does not make sense here. Thus, as mentioned above, the absence of self-loops indicates that nothing is happening at the moment, and the MC is just waiting according to an exponential distribution.

Example 16.4

Consider the following CTMC:



You wait in state i for $\text{Exp}(q_i)$ time, where $q_i = q_{ij} + q_{ik}$, then jump to state j with probability $P_{ij} = \frac{q_{ij}}{q_i}$ or k with probability $P_{ik} = \frac{q_{ik}}{q_i}$. Since the rows of $\mathbf{Q}$ sum to 0, $P_{ij} + P_{ik} = 1$.

17 Lecture 17

17.1 CTMCs Cont.

Definition 17.1: Jump Rates
Let $q_{ij} = q_i P_{ij}$ for all $i, j \in S$ be the "jump rates" of a CTMC.

Another point of view for the working of a CTMC: on entering state i, consider independent random variables $T_i \sim \text{Exp}(q_{ij})$ for $j \in S \setminus \{i\}$. Then, the CTMC jumps to state $j^* = \text{argmin}_{j \in S}(T_j : j \in S)$ after time T_{j^*} has elapsed. This is equivalent to setting random timers, modeled using exponential RVs with specific parameters, for all of your adjacent states and transitioning to the one whose timer goes off first.

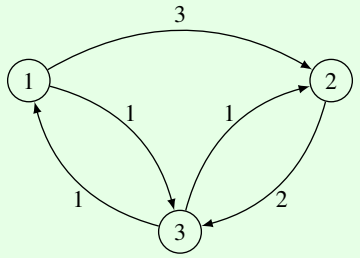
Example 17.1
If

$$\mathbf{Q} = \begin{bmatrix} -4 & 3 & 1 \\ 0 & -2 & 2 \\ 1 & 1 & -2 \end{bmatrix}$$

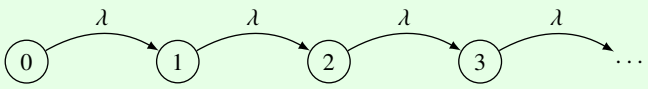
The transition rates will be $q_1 = 4$, $q_2 = 2$ and $q_3 = 2$. Then, the jump chain will have a transition matrix given by

$$\mathbf{P} = \begin{bmatrix} 0 & 3/4 & 1/4 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix}$$

For CTMCs, when drawing a state transition diagram, we label the transitions with their jump rates:



Example 17.2
Consider a Poisson process with rate λ:

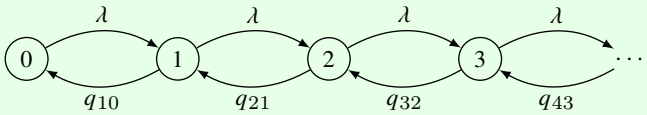


Then,

$$q_{n,n+1} = \lambda \quad n \geq 0$$
$$q_{ij} = 0 \quad \text{otherwise}$$

Example 17.3: M/M/S Queue
Customers arrive to a system with S servers according to a $\text{PP}(\lambda)$. If a server is available, the arrival immediately enters service. Service times are $\sim_{\text{IID}} \text{Exp}(\mu)$. If no server is available, the arrival waits until one becomes available. Let $(X_t)_{t \geq 0}$ denote the number of customers in the system at time $t \geq 0$ [in system means in queue and in service]. Model this as a CTMC, i.e., specify the jump rates. Let the state space be $\{0, 1, 2, \dots\}$. Then the state transition diagram will

look like the following:



$$q_{n,n+1} = \lambda$$
$$q_{n,n-1} = \begin{cases} n\mu & 1 \le n \le S \\ S\mu & n > S \end{cases}$$

For $1 \le n \le S$, the number of customers in then queue go down after one of the n servers is done being used. Since server times are exponential, this time is modeled using the minimum of n IID exponential RVs. Similarly, for $n > S$, all of the servers are busy, so the customer queue only decreases after one of the S servers free up. Just like before, this waiting time is then modeled using the minimum of S IID exponential RVs.

Example 17.4: Birth Death Chain

Individuals give birth $\sim \text{PP}(\lambda)$ and have lifetimes of IID duration $\text{Exp}(\mu)$. Let X_t be the number of individuals at time t. The state space is the set of naturals. The jump rates are,

$$q_{n,n+1} = n\lambda$$
$$q_{n,n-1} = n\mu$$

At state n, there are n individuals that can give birth, so the waiting time until the next birth is the minimum of n $\text{Exp}(\lambda)$ RVs. Similarly, there are also n individuals that can die, so the waiting time before one dies is the minimum of n $\text{Exp}(\mu)$ RVs. Thus, $q_n = n(\lambda + \mu)$ so the discrete time jump-chain is also a birth death chain with

$$P_{n,n+1} = \frac{\lambda}{\lambda + \mu}$$
$$P_{n,n-1} = \frac{\mu}{\lambda + \mu}$$

17.2 Stationary Distributions of CTMCs

Definition 17.2: Stationary Distribution

A probability vector $\boldsymbol{\pi}$ is a stationary distribution for a CTMC with rate matrix $\mathbf{Q}$ iff

$$\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$$

This is called the "rate conservation principle".

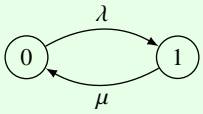
Expanding out the definition above,

$$\pi_j q_j = \sum_{i \in S} \pi_i q_{ij}$$

where $\pi_j q_j$ is the rate at which transitions are made out of j and $\pi_i q_{ij}$ is the rate at which transitions are made into state j from state i (assuming $\mathbb{P}[X_t = i] = \pi_i$).

Example 17.5

Consider the following simple 2-state CTMC:



Then,

$$\mathbf{Q} = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$$
$$\boldsymbol{\pi} = \begin{bmatrix} \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \end{bmatrix}$$

Just like DTMCs, there is classification of states in CTMCS too:

- $i \leftrightarrow j$ in CTMC $\iff$ $i \leftrightarrow j$ in jump chain
- Classes in a CTMC are the same as those in its associated jump chain.
- State j is transient if, given $X_0 = j$, $(X_t)_{t\geq 0}$ re-enters state j finitely many times with probability 1. State j is recurrent otherwise.
- For a recurrent state j, define
$$T_j = \min\{t \geq 0 : X_t = j \text{ and } X_s \neq j \text{ for some } s < t\}$$
This is the time of the first re-entry back into state j.
- State j is positive recurrent if $\mathbb{E}\left[T_j \mid X_0 = j\right] < \infty$ and null recurrent if $\mathbb{E}\left[T_j \mid X_0 = j\right] = \infty$
- There is no concept of periodicity in CTMCs because any state can visit any accessible state in any amount of time with positive probability.
- Transience and positive/null recurrence are still class properties.

Theorem 17.1: Big Theorem

This theorem characterizes the long-term behavior of CTMCs, just like we did for DTMCs. Define

$$P_{ij}^t = \mathbb{P}[X_t = j | X_0 = i]$$
$$m_j = \mathbb{E}\left[T_j | X_0 = j\right]$$

Let $(X_t)_{t\geq 0}$ be an irreducible MC. Exactly one of the following true:

1. Either all states are transient or all are null recurrent. In this case, no stationary distribution exists and
$$\lim_{t\to\infty} P_{ij}^t = 0, \forall i, j \in S$$
2. All states are positive recurrent. In this case, a stationary distribution π exists. It is unique and satisfies
$$\pi_j = \frac{1}{m_j q_j} = \lim_{t\to\infty} P_{ij}^t, \forall i, j \in S$$

The SD of a CTMC, given by $\boldsymbol{\pi}$, is not the same as the SD of its jump chain, given by $\tilde{\boldsymbol{\pi}}$. In particular,

$$\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$$
$$\pi_i q_i = \sum_{j\neq i} \pi_j Q_{ji}$$
$$= \sum_{j\neq i} \pi_j P_{ji} q_j$$

$$= \sum_j (\pi_j q_j) P_{ji} \quad (\text{since } P_{ii} = 0)$$

The vector $(\pi_i q_i)_{i \in S}$ is an eigenvector of the jump chain transition matrix $\mathbf{P}$. Thus, the stationary distribution of the jump chain is just this vector normalized:

$$\tilde{\pi}_i = \frac{\pi_i q_i}{\sum_j \pi_j q_j}$$

Similarly,

$$\pi_i = \frac{\frac{\tilde{\pi}}{q_i}}{\sum_j \frac{\tilde{\pi}}{q_j}}$$

Example 17.6: M/M/∞ Queue

$$q_{n,n+1} = \lambda \quad \text{arrivals} \sim \text{PP}(\lambda)$$
$$q_{n,n-1} = n\mu \quad \text{service times are the min of IID } \text{Exp}(\mu)$$

Solve $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$ to get $\pi_n = e^{-\frac{\lambda}{\mu}} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}$. By the Big Theorem, $X_t \xrightarrow{d} \text{Poisson}(\frac{\lambda}{\mu})$ where X_t is the number of people in the system at time t.

18 Lecture 18

18.1 First Step Analysis

Exactly the same idea as for DTMCs. In fact, hitting probabilities is exactly the same (but done on the jump chain). Only difference is when we consider time-dependent quantities like expected hitting time.

If $A \subseteq S$, define $T_A = \min\{t \geq 0 : X_t \in A\}$. To compute $\mathbb{E}[T_A \mid X_0 = j]$, use the same strategy as for DTMCs, except account for holding times. Define

$$t_i = \mathbb{E}[T_A \mid X_0 = i]$$
$$t_i = 0, \forall i \in A$$

For $i \notin A$, use the law of total expectation

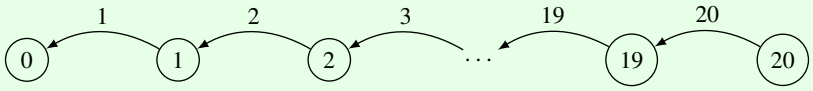
$$\underbrace{\mathbb{E}[T_A \mid X_0 = i]}_{t_i} = \mathbb{E}[\text{time to hold in state } i] + \sum_{j \in S} P_{ij}\mathbb{E}[T_A \mid X_0 = j]$$

where P_{ij} are the transition probabilities in its associated jump chain. Summarizing, we obtain our first step equations

$$\text{FSE} = \begin{cases} t_i = \frac{1}{q_i} + \sum_j P_{ij} t_j & i \notin A \\ t_i = 0 & i \in A \end{cases}$$

Example 18.1

There are 20 lightbulbs with lifetimes $\sim_{\text{IID}} \text{Exp}(1)$. Assuming they are all on at time $t = 0$, how long does it take until they all burn out? Consider representing this process with the following CTMC:



The first step equations are

$$t_n = \mathbb{E}[T_0 \mid X_0 = n]$$
$$t_0 = 0$$
$$t_1 = 1 + t_0 = 1$$
$$t_2 = \frac{1}{2} + t_1 = \frac{1}{2} + 1$$
$$t_3 = \frac{1}{3} + t_2 = \frac{1}{3} + \frac{1}{2} + 1$$
$$\vdots$$
$$t_{20} = 1 + \frac{1}{2} + \cdots + \frac{1}{20} \approx 3.6$$

Observe that $\mathbb{E}[\text{time in state } i]$ is the expectation of the minimum of i exponentials.

18.2 Uniformization

For some context, consider a CTMC with transition rates $(q_i)_{i \in S}$, and assume there exists an $M > 0$ such that $q_i \leq M, \forall i \in S$. Let $\mathbf{P}^t$ denote the matrix of transition probabilities at time $t \geq 0$, i.e., $\mathbf{P}^t_{ij} = \mathbb{P}[X_t = j \mid X_0 = i]$. Markovity gives us $\mathbf{P}^{s+t} = \mathbf{P}^s\mathbf{P}^t, \forall s, t \geq 0$ (Chapman-Kolmogorov Equations). We can show that $\mathbf{P}^h \approx \mathbf{I} + h\mathbf{Q} + \mathbf{o}(h)$ for $h \approx 0$. So,

$$\begin{aligned} \mathbf{P}^{t+h} &= \mathbf{P}^t\mathbf{P}^h \\ &= \mathbf{P}^t(\mathbf{I} + h\mathbf{Q} + \mathbf{o}(h)) \end{aligned}$$

$$\implies \frac{\mathbf{P}^{t+h} - \mathbf{P}^t}{h} = \mathbf{P}^t\mathbf{Q} + \frac{\mathbf{o}(h)}{h}$$

Letting $h \downarrow 0$,

$$\frac{\partial}{\partial t}\mathbf{P}^t = \mathbf{P}^t\mathbf{Q}$$

This is called the Kolmogorov Forward Equation. In particular, this differential equation has a unique solution given by

$$\mathbf{P}^t = e^{t\mathbf{Q}} = \sum_{k\geq 0} \frac{(t\mathbf{Q})^k}{k!}, \forall t \geq 0$$

How to compute P^t for large state space? This is where uniformization comes in.

Definition 18.1: Uniformization

Take $\gamma \geq M \geq q_{ij}, \forall i, j \in S$. Then, define a *uniformized* DTMC with transition probabilities

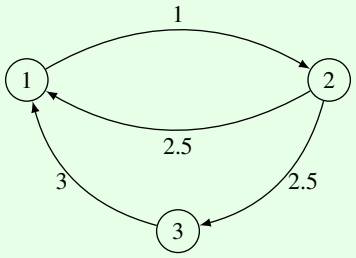
$$P_{ij} = \frac{q_{ij}}{\gamma}$$
$$P_{ii} = 1 - \frac{q_i}{\gamma}$$

Note 18.1

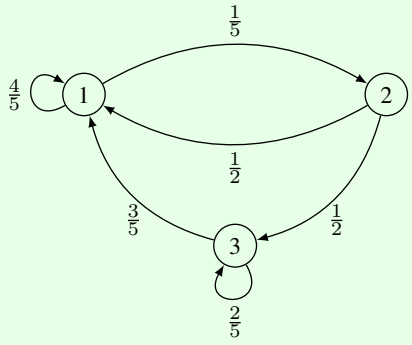
These are not the transition probabilities of the corresponding jump chain.

Example 18.2

Consider the following CTMC:



Letting $\gamma = 5$, we get the following uniformized DTMC:



If $\mathbf{P}_u$ is the transition matrix for the uniformized DTMC, then by its construction,

$$\mathbf{P}_u = \mathbf{I} + \frac{1}{\gamma}\mathbf{Q}$$

So observe

$$\boldsymbol{\pi}\mathbf{P}_u = \boldsymbol{\pi} + \frac{1}{\gamma}\boldsymbol{\pi}\mathbf{Q}$$

Then, $\pi \mathbf{P}_u = \pi \iff \pi \mathbf{Q} = \mathbf{0} \iff \pi$ is a stationary distribution for both the CTMC and the uniformized DTMC. The point of uniformization is to (approximately) compute $\mathbf{P}^t$ by running the uniformized DTMC for some number of steps. How does it work? Note that

$$\mathbf{P}_u^n = \left(\mathbf{I} + \frac{1}{\gamma}\mathbf{Q}\right)^n \approx e^{\frac{n}{\gamma}\mathbf{Q}}$$

where $\mathbf{P}_u^n$ is the n-step transition probability matrix for uniformized DTMC (following the Kolmogorov-Chapman equations). Therefore, to estimate $\mathbf{P}^t$, run the uniformized chain for $n \approx \gamma t$ steps (because $\mathbf{P}^t = e^{t\mathbf{Q}} \approx e^{\frac{n}{\gamma}\mathbf{Q}} = \mathbf{P}_u^n$).

In summary, Euler schemes are discrete-time approximations to solving differential equations. Uniformization is an Euler scheme approach to compute continuous time transition probabilities by simulating a DTMC.

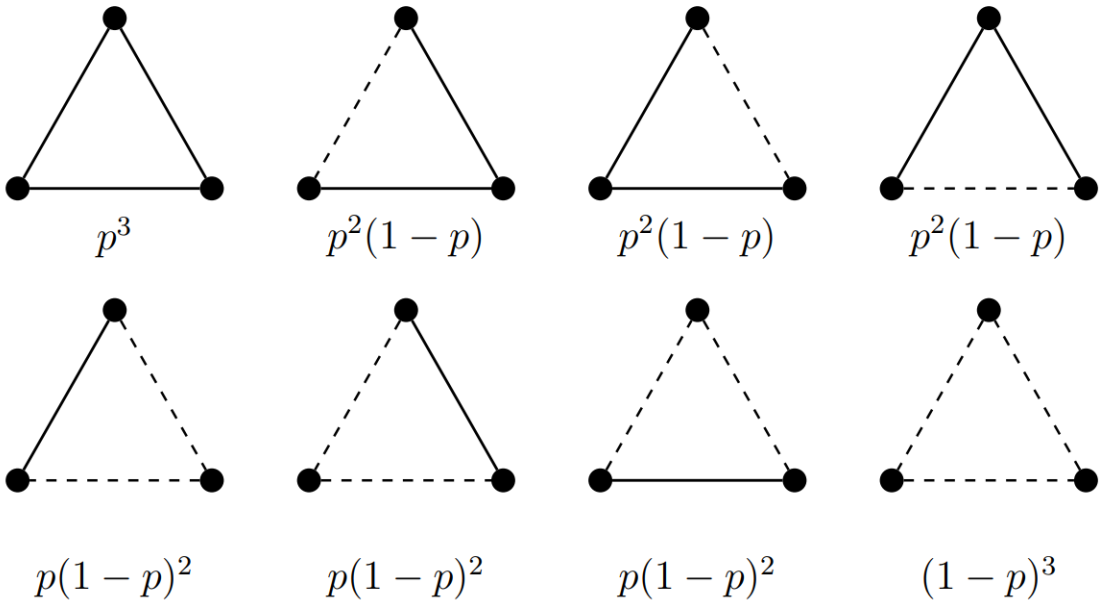
18.3 Random Graphs

A lot of objects in EECS and beyond are modeled by graphs (social networks, dependency structure in databases/programs, tournaments, epidemiology, etc.). The simplest class of random graphs is given by the Erdős-Rényi ensemble (i.e., the IID coin flips of the graph world).

Definition 18.2: Erdős-Rényi Random Graph

Fix $n \geq 1$ and $p \in [0,1]$. A random graph $\mathcal{G}(n,p)$ is an undirected graph on n vertices, where each edge is placed independently with probability p.

For $n = 3$, the different possible ER graphs and their respective probabilities are listed below



What are some of the types of questions that we can ask? If $n \to \infty$, how should p scale with n so that a random graph has a certain property P with high probability? Like for capacity (in information theory), there is often a sharp "threshold" behavior that is also observable here.

Theorem 18.1: Friedgut and Kalai, 1996

Every "monotone" graph property P has a sharp threshold t_n such that

- $p \gg t_n \implies \mathcal{G}(n,p)$ has P with high probability
- $p \ll t_n \implies \mathcal{G}(n,p)$ does not have P with high probability

Example 18.3

Different properties and their thresholds:

1.
   - If $p \ll \frac{1}{n^2}$, then there are no edges in $\mathcal{G}(n,p)$ with high probability (use Markov's inequality)
   - If $p \gg \frac{1}{n^2}$, then $\mathcal{G}(n,p)$ has edges with high probability
2.
   - If $p \ll \frac{1}{n}$, then there is no cycle with high probability
   - If $p \gg \frac{1}{n}$, then there is a cycle with high probability
3.
   - If $p \ll \frac{1}{n}$, then the largest connected component is of size $\mathcal{O}(\log n)$ with high probability
   - If $p \gg \frac{1}{n}$, then the largest connected component is of size $\Theta(n)$

19 Lecture 19

19.1 Random Graphs Cont.

How to interpret the thresholds? If p is much larger than the threshold, then a given property holds. If it's much smaller, then the said property doesn't hold. Usually, the "much more/less" qualification depends on the problem to some extent.

Let's be explicit for the property of having edges. Let X be the number of edges in $G(n, p_n)$, and let's take $p_n = \frac{c}{n^2}$. Note that $X \sim \text{Binomial}\left(\binom{n}{2}, p_n\right)$. Then,

$$\mathbb{P}[X = 0] = (1 - p_n)^{\binom{n}{2}} \to e^{-\frac{c}{2}}$$

So, if $c \gg 1$, then $\mathbb{P}[X = 0] \approx 0$, and if $c \ll 1$, then $\mathbb{P}[X = 0] \approx 1$.

19.2 Connectivity Threshold

Recall that a graph is "connected" if there exists a path between any given pair of vertices. We saw in the previous lecture that the threshold for emergence of a "giant component" (connected subgraph of size $\Theta(n)$) was $\frac{1}{n}$. The connectivity threshold is about $\approx \log n$ times larger.

Theorem 19.1: Erdős-Rényi
Fix $\lambda > 0$, and let $p_n = \lambda \frac{\log n}{n}$. Then,

- If $\lambda > 1$, then $\mathbb{P}[G(n, p_n) \text{ is connected}] \to 1$
- If $\lambda < 1$, then $\mathbb{P}[G(n, p_n) \text{ is connected}] \to 0$

Quick lemma needed for the overall proof: if X is a random variable, then $\mathbb{P}[X = 0] \leq \frac{\text{Var}(X)}{(\mathbb{E}[X])^2}$.

Proof: Apply the law of total expectation to the definition of variance:

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&= \mathbb{P}[X = 0]\mathbb{E}\left[(X - \mathbb{E}[X])^2 \mid X = 0\right] + \mathbb{P}[X \neq 0]\mathbb{E}\left[(X - \mathbb{E}[X])^2 \mid X \neq 0\right] \\
&\geq \mathbb{P}[X = 0]\,(\mathbb{E}[X])^2
\end{aligned}$$

Proof: Proof of the connectivity threshold:

- Case $\lambda < 1$: We will show something stronger: $\mathbb{P}[\mathcal{G}(n, p) \text{ is disconnected}] \geq \mathbb{P}[\mathcal{G}(n, p) \text{ contains isolated vertices}] \to 1$.
  Let X be the number of isolated vertices. Use the lemma above to show that $\mathbb{P}[X = 0] \to 0$ or equivalently $\mathbb{P}[X \geq 1] \to 1$. Let I_i be the indicator that vertex i is isolated. Then,

$$\begin{aligned}
X &= \sum_{i=1}^{n} I_i \\
\mathbb{E}[X] &= n\mathbb{E}[I_1] \\
&= n\mathbb{P}[\text{vertex 1 is isolated}] \\
&= n \underbrace{(1 - p)^{n-1}}_{q}
\end{aligned}$$

  The expression above follows since a given vertex is not connected to a different vertex (chosen among any of the remaining vertices) independently with probability $1 - p$. Moreover,

$$\begin{aligned}
\text{Var}(X) &= \sum_i \text{Var}(I_i) + \sum_{i \neq j} \text{Cov}(I_i, I_j) \\
&= nq(1 - q) + n(n - 1)\frac{pq^2}{1 - p}
\end{aligned}$$

The above follows since $I_i \sim \text{Bernoulli}(q)$. Using the lemma,

$$\begin{aligned}\mathbb{P}[X=0] &\leq \frac{nq(1-q)+n(n-1)\frac{pq^2}{1-p}}{n^2q^2} \\ &= \frac{1-q}{nq} + \frac{n-1}{n}\cdot\frac{p}{1-p} \\ &\leq \frac{1-q}{nq} + \frac{p}{1-p}\end{aligned}$$

As $n \to \infty$, note that $p = \lambda\frac{\log n}{n} \to 0$. Thus, $\frac{p}{1-p} \to 0$. Moreover, observe that,

$$\begin{aligned}\log(nq) &= \log\Big(n(1-p)^{n-1}\Big) \\ &= \log(n) + (n-1)\log(1-p) \\ &\approx \log(n) - (n-1)p \\ &= \log(n) - (n-1)\lambda\frac{\log n}{n} \\ &\approx \log(n) - \lambda\log(n) \\ &= \log\Big(n^{1-\lambda}\Big) \\ &\to \infty\end{aligned}$$

Thus, $nq \to \infty$ as $n \to \infty$, and $\frac{1}{nq} \to 0$. Then, $\mathbb{P}[X=0] \to 0$.

- Case $\lambda > 1$:

$$\begin{aligned}\mathbb{P}[\mathcal{G}(n,p) \text{ disconnected}] &= \mathbb{P}\left[\bigcup_{k=1}^{\frac{n}{2}} \{\exists \text{ set of } k \text{ disconnected vertices}\}\right] \\ &\leq \sum_{k=1}^{\frac{n}{2}} \mathbb{P}[\exists \text{ set of } k \text{ disconnected vertices}] \\ &\leq \sum_{k=1}^{\frac{n}{2}} \binom{n}{k} \mathbb{P}[\text{specific set of } k \text{ vertices disconnected from rest}] \\ &= \sum_{k=1}^{\frac{n}{2}} \binom{n}{k} (1-p)^{k(n-k)} \\ &\to 0\end{aligned}$$

for $\lambda > 1$. The above expression follows because if there is a set of k vertices, then it is disconnected from the remaining $n-k$ vertex graph through the absence of the $k \times (n-k)$ edges between them.

19.3 Statistical Inference

Goal of statistical hypothesis testing is to use probabilistic insight and reasoning to design procedures to accomplish specific tasks. The basic setup for hypothesis testing goes as follows

$$\underbrace{X}_{\text{state of nature}} \longrightarrow \underbrace{P_{Y|X}}_{\text{model}} \longrightarrow \underbrace{Y}_{\text{observation}}$$

- X takes values in $\{0, 1, \ldots, M-1\}$ (so M hypotheses to consider)
- The model generates "likelihoods" $P_{Y|X}$ (which can be a PDF $f_{Y|X}(\cdot \mid x)$ or a PMF $p_{Y|X}(\cdot \mid x)$) of observing Y given some specific X

The variable X may or may not be a random variable (i.e., it may not be described by a known probability distribution).

Definition 19.1: Prior
When X is a random variable with known distribution such that $P(X = i) = \pi_i$, for $i = 0, \ldots, M-1$, then we call π a "prior" on X. This concept is the foundation for Bayesian inference.

Example 19.1
Let $X \in \{\text{Healthy}, \text{Covid}, \text{Flu}\}$ and Y be symptoms observed [Assumption: we are always given the likelihood model $P_{Y|X}$]. If, for example,

$$\mathbb{P}[X = H] = 0.9$$
$$\mathbb{P}[X = F] = 0.03$$
$$\mathbb{P}[X = C] = 0.07$$

then this is a prior such that

$$\pi_H = 0.9$$
$$\pi_F = 0.03$$
$$\pi_C = 0.07$$

This reflects our prior knowledge of the state of nature (ex. prevalence of flu/covid in several populations).

By Bayes rule, if we observe $Y = y$, then the "a posteriori" probability of $X = x$ is given by

$$\mathbb{P}[X = x \mid Y = y] = \frac{P_{Y|X}(y \mid x)\pi_x}{\sum_{\tilde{x}} P_{Y|X}(y \mid \tilde{x})\pi_{\tilde{x}}}$$

Note that the denominator does not depend on x, only on y. Think of the above as an update to the prior given some observations. So, this motivates the Maximum A Posteriori (MAP) estimate.

Definition 19.2: MAP
The most likely x after observing $Y = y$ is given by

$$\begin{aligned}\hat{X}_{\text{MAP}}(y) &= \operatorname*{argmax}_x P_{X|Y}(x \mid y) \\ &= \operatorname*{argmax}_x P_{Y|X}(y \mid x)\pi_x\end{aligned}$$

A MAP estimate depends on likelihoods and priors. Moreover, for any other $\hat{X}(y)$ other than $\hat{X}_{\text{MAP}}(y)$,

$$\mathbb{P}\left[\hat{X}(y) \neq X \mid Y = y\right] \geq \mathbb{P}\left[\hat{X}_{\text{MAP}}(y) \neq X \mid Y = y\right]$$

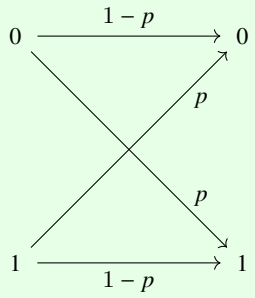
so the MAP estimate minimizes the probability of error.

What if we don't have a prior? One simple strategy is to assume that π is uniform over all x. In this case, the MAP estimate reduces to maximizing the likelihood of the observation over the hypotheses.

Definition 19.3: MLE
Assuming a uniform prior, the Maximum Likelihood Estimate (MLE) can be defined as

$$\hat{X}_{MLE}(y) = \operatorname*{argmax}_x P_{Y|X}(y \mid x)$$

Example 19.2
Consider the BSC(p) channel given below such that X is the channel input and Y is the channel output



The MLE estimate of X is given by

$$\hat{X}_{\text{MLE}}(y) = \operatorname*{argmax}_x P_{Y|X}(y \mid x) = \begin{cases} y & \text{if } p \leq \frac{1}{2} \\ 1-y & \text{if } p \geq \frac{1}{2} \end{cases}$$

The MAP estimate of X is given by

- For $y = 0$,

$$P_{Y|X}(0 \mid x)\pi(x) = \begin{cases} (1-p)\pi_0 & x = 0 \\ p(1-\pi_0) & x = 1 \end{cases}$$

$$\hat{X}_{\text{MAP}}(0) = \begin{cases} 0 & \text{if } p < \pi_0 \\ 1 & \text{if } p \geq \pi_0 \end{cases}$$

  Note that we choose $x = 0$ when

$$\begin{aligned} (1-p)\pi_0 &> p(1-\pi_0) \\ \pi_0 - p\pi_0 &> p - p\pi_0 \\ p &< \pi_0 \end{aligned}$$

- For $y = 1$,

$$P_{Y|X}(1 \mid x)\pi(x) = \begin{cases} p\pi_0 & x = 0 \\ (1-p)(1-\pi_0) & x = 1 \end{cases}$$

$$\hat{X}_{\text{MAP}}(1) = \begin{cases} 0 & \text{if } 1-p < \pi_0 \\ 1 & \text{if } 1-p \geq \pi_0 \end{cases}$$

20 Lecture 20

20.1 Statistical Inference Cont.

Definition 20.1: Likelihood Ratio
In the problem of binary hypothesis testing (i.e., $M = 2$), we can define the likelihood ratio between two hypotheses as

$$L(y) = \frac{P_{Y|X}(y \mid 1)}{P_{Y|X}(y \mid 0)}$$

This allows us to reformulate the BSC example from before as follows

$$\hat{X}_{\text{MLE}}(y) = \begin{cases} 1 & \text{if } L(y) \geq 1 \\ 0 & \text{if } L(y) < 1 \end{cases}$$

$$\hat{X}_{\text{MAP}}(y) = \begin{cases} 1 & \text{if } L(y) \geq \frac{\pi_0}{\pi_1} \\ 0 & \text{if } L(y) < \frac{\pi_0}{\pi_1} \end{cases}$$

Both $\hat{X}_{\text{MLE}}(y)$ and $\hat{X}_{\text{MAP}}(y)$ can be expressed as a "threshold test" where I just threshold the likelihood ratio evaluated for my observations! This was just a discrete example. Let's look at a case of continuous observations below.

Example 20.1
Let $X \in \{0, 1\}$ and $Y = X + Z$ for $Z \sim \mathcal{N}(0, \sigma^2)$ independent of X. The individual likelihoods are

$$f_{Y|X}(y \mid 0) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}$$

$$f_{Y|X}(y \mid 1) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-1)^2}{2\sigma^2}}$$

The likelihood ratio is

$$\begin{aligned} L(y) &= \frac{f_{Y|X}(y \mid 1)}{f_{Y|X}(y \mid 0)} \\ &= e^{\frac{y}{\sigma^2} - \frac{1}{2\sigma^2}} \end{aligned}$$

The two estimates are

$$\hat{X}_{\text{MAP}}(y) = \begin{cases} 1 & \text{if } L(y) \geq \frac{\pi_0}{\pi_1} \\ 0 & \text{if } L(y) < \frac{\pi_0}{\pi_1} \end{cases}$$

$$\hat{X}_{\text{MLE}}(y) = \begin{cases} 1 & \text{if } L(y) \geq 1 \\ 0 & \text{if } L(y) < 1 \end{cases}$$

Where did the $\frac{\pi_0}{\pi_1}$ threshold come from? Recall from the definition of MAP that we choose $X = 1$ if

$$\begin{aligned} P_{Y|X}(y \mid 1)\pi_1 &\geq P_{Y|X}(y \mid 0)\pi_0 \\ \underbrace{\frac{P_{Y|X}(y \mid 1)}{P_{Y|X}(y \mid 0)}}_{L(y)} &\geq \frac{\pi_0}{\pi_1} \end{aligned}$$

20.2 Binary Hypothesis Testing

Both of the previous examples are instances of "binary hypothesis testing", i.e., there are only two hypotheses to discriminate between given our observation y:

$$H_0 : Y \sim P_{Y|X=0} \quad \text{(the null hypothesis)}$$

$$H_1 : Y \sim P_{Y|X=1} \quad \text{(the alternate hypothesis)}$$

To choose one of the hypotheses, we can define a decision rule/test $\hat{X} : \mathcal{Y} \to \{0, 1\}$. There are two fundamental types of error associated with any binary test, namely the

- Type I Error (False Positive) Probability: $\mathbb{P}\left[\hat{X}(Y) = 1 \mid X = 0\right]$
- Type II Error (False Negative) Probability: $\mathbb{P}\left[\hat{X}(Y) = 0 \mid X = 1\right]$

Note 20.1
Another name for the type I error probability is the probability of false alarm (PFA). Similarly, you can define the probability of correct detection (PCD) as $\mathbb{P}\left[\hat{X}(Y) = 1 \mid X = 1\right]$ and denote the type 2 error probability by 1 - PCD. These names are slightly more descriptive and easier to remember.

The goal of binary hypothesis testing is to choose a test that minimizes the type II error probability (maximizes the probability of correct detection) subject to a constraint on the type I error probability (the probability of false alarm). That is, for some $\beta \geq 0$, find

$$\hat{X}^*_\beta = \underset{\hat{X}:\mathbb{P}\left[\hat{X}(y)=1|X=0\right]\leq\beta}{\operatorname{argmin}} \mathbb{P}\left[\hat{X}(y) = 0 \mid X = 1\right] = \underset{\hat{X}:\mathbb{P}\left[\hat{X}(y)=1|X=0\right]\leq\beta}{\operatorname{argmax}} \mathbb{P}\left[\hat{X}(y) = 1 \mid X = 1\right]$$

Theorem 20.1: Neyman-Pearson Lemma
Given $\beta \in [0, 1]$, the most optimal decision rule is a (randomized) threshold test defined as follows

$$\hat{X}_\beta(y) = \begin{cases} 0 & \text{if } L(y) < \lambda \\ \text{Bernoulli}(\gamma) & \text{if } L(y) = \lambda \\ 1 & \text{if } L(y) > \lambda \end{cases}$$

where λ, γ are chosen to set $\mathbb{P}\left[\hat{X}(y) = 1 \mid X = 0\right] = \beta$.

This test $\hat{X}^*_\beta$ is known as the "Neyman-Pearson Rule." It is the most powerful test that minimizes type II error probability subject to the constraint that $\mathbb{P}[\text{type I error}] \leq \beta$. We will prove its optimality next time.

Note 20.2
If the likelihood $L(y)$ is monotonically increasing, then you can convert the optimal test to the following form:

$$\hat{X}_\beta(y) = \begin{cases} 0 & \text{if } y < \lambda \\ \text{Bernoulli}(\gamma) & \text{if } y = \lambda \\ 1 & \text{if } y > \lambda \end{cases}$$

If its monotonically decreasing, just flip the inequalities above.

Example 20.2
Consider the Gaussian example from before where

$$X = 0 : Y \sim \mathcal{N}(0, \sigma^2)$$
$$X = 1 : Y \sim \mathcal{N}(1, \sigma^2)$$
$$L(y) = e^{\frac{y}{\sigma^2} - \frac{1}{2\sigma^2}}$$

Let's say that we want to bound the type I error Probability by β. Then, Neyman-Pearson tells us to only consider threshold tests, and we just need to choose the λ, γ parameters appropriately. Note that $\mathbb{P}[L(Y) = \lambda] = 0$ for any λ because Y, and hence $L(Y)$, is a continuous random variable. So, no randomization to worry about in this case and we can effectively ignore γ.

How about λ? Set

$$\begin{aligned}
\beta &= \mathbb{P}\left[\hat{X}(Y) = 1 \mid X = 0\right] \\
&= \mathbb{P}[L(Y) \geq \lambda \mid X = 0] \\
&= \mathbb{P}\left[\exp\left(\frac{Y}{\sigma^2} - \frac{1}{2\sigma^2}\right) \geq \lambda \mid X = 0\right] \\
&= \mathbb{P}\left[Y \geq \frac{1}{2} + \sigma^2 \log(\lambda) \mid X = 0\right] \\
&= \mathbb{P}\left[\underbrace{\frac{Y}{\sigma}}_{\mathcal{N}(0,1)} \geq \frac{1}{2\sigma} + \sigma \log(\lambda) \mid X = 0\right] \\
&= 1 - \Phi\left(\frac{1}{2\sigma} + \sigma \log(\lambda)\right)
\end{aligned}$$

Solve for λ in terms of β, σ.

21 Lecture 21

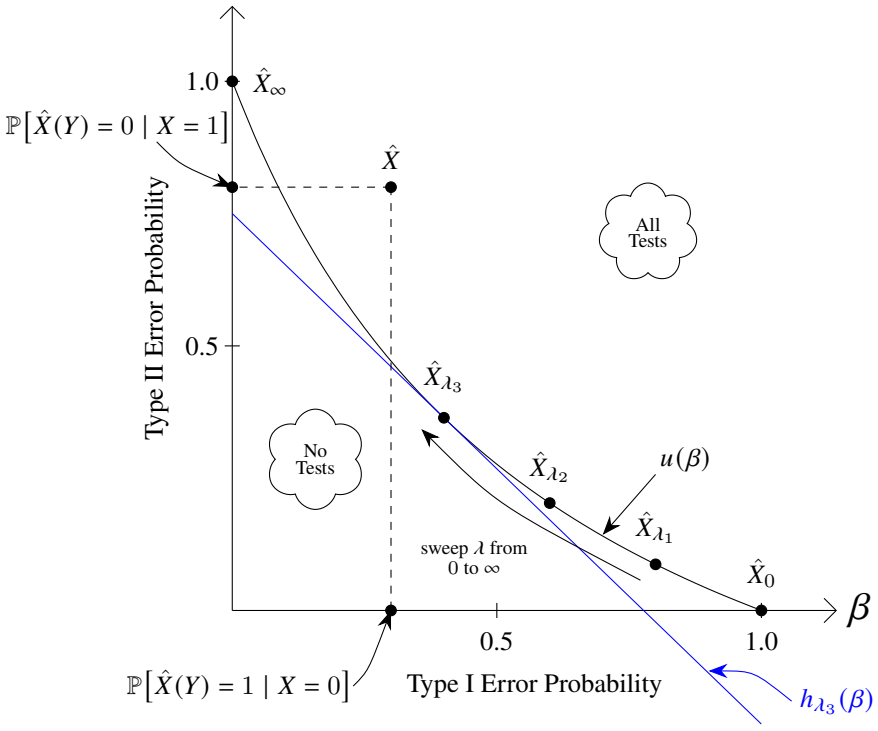
21.1 Neyman-Pearson Lemma Proof

Proof: The two things that we care about for a test are

$$\begin{aligned}\hat{X} : Y \to \hat{X}(y) \in \{0,1\}\\ \text{Type I error rate } &= \mathbb{P}\big[\hat{X}(Y) = 1 \mid X = 0\big]\\ \text{Type II error rate } &= \mathbb{P}\big[\hat{X}(Y) = 0 \mid X = 1\big]\end{aligned}$$

Let the threshold test with threshold λ be denoted by $\hat{X}_\lambda$. For example,

- $\hat{X}_0$ always returns 1
- $\hat{X}_\infty$ always returns 0



Let $u(\beta)$ be known as an "error curve" and define it as

$$u(\beta) = \max_{\lambda \geq 0} \underbrace{\mathbb{P}\big[\hat{X}_\lambda(Y) = 0 \mid X = 1\big] + \lambda\Big(\mathbb{P}\big[\hat{X}_\lambda(Y) = 1 \mid X = 0\big] - \beta\Big)}_{h_\lambda(\beta)}$$

For those who have taken EECS 127, note that $u(\beta)$ is convex since it is a pointwise maximum of affine functions. The goal of this proof is to show that all thresholds tests lie on the error curve and that every other test lies above it. First, observe that for a fixed λ_0,

$$u\Big(\mathbb{P}\big[\hat{X}_{\lambda_0}(Y) = 1 \mid X = 0\big]\Big) \geq \underbrace{\mathbb{P}\big[\hat{X}_{\lambda_0}(Y) = 0 \mid X = 1\big]}_{\text{Type II Error prob. for } \hat{X}_\lambda}$$

from the definition of u (choosing $\lambda = \lambda_0$ will at least lower bound it by the RHS). Thus, $\hat{X}_{\lambda_0}$ lies on or below the error curve. Since this bound holds for all λ_0, this implies that all threshold tests lie on or below the error curve.

Now, we will show that all tests (both threshold and non-threshold) lie above the error curve. Fix $\lambda \in [0, \infty]$. Impose an artificial prior on X with $\frac{\pi_0}{\pi_1} = \lambda$. In this case, let $\hat{X}_{\text{MAP}}(Y) = \hat{X}_\lambda(Y)$. The MAP test has a special property in that it minimizes the probability of error, i.e.,

$$\mathbb{P}\left[\hat{X}_{\text{MAP}}(Y) \neq X\right] \leq \mathbb{P}\left[\hat{X} \neq X\right]$$

for any $\hat{X}$. Equivalently, using the law of total probability,

$$\pi_0 \mathbb{P}\left[\hat{X}(Y) = 1 \mid X = 0\right] + \pi_1 \mathbb{P}\left[\hat{X}(Y) = 0 \mid X = 1\right] \geq \pi_0 \mathbb{P}\left[\hat{X}_\lambda(Y) = 1 \mid X = 0\right] + \pi_1 \mathbb{P}\left[\hat{X}_\lambda(Y) = 1 \mid X = 0\right]$$
$$\lambda \mathbb{P}\left[\hat{X}(Y) = 1 \mid X = 0\right] + \mathbb{P}\left[\hat{X}(Y) = 0 \mid X = 1\right] \geq \lambda \mathbb{P}\left[\hat{X}_\lambda(Y) = 1 \mid X = 0\right] + \mathbb{P}\left[\hat{X}_\lambda(Y) = 1 \mid X = 0\right]$$

Rearranging this expression,

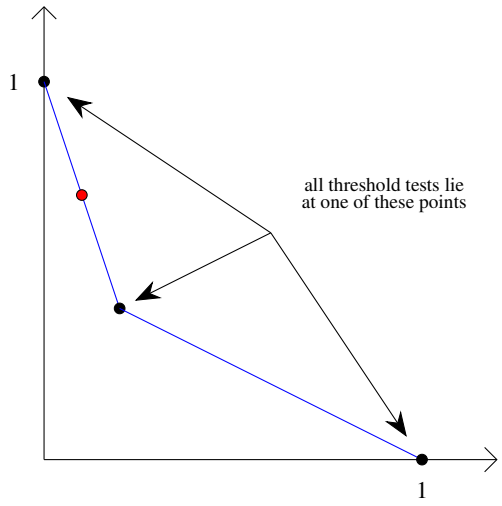
$$\mathbb{P}\left[\hat{X}(Y) = 0 \mid X = 1\right] \geq \mathbb{P}\left[\hat{X}_\lambda(Y) = 0 \mid X = 1\right] + \lambda\Big(\mathbb{P}\left[\hat{X}_\lambda(Y) = 1 \mid X = 0\right] - \mathbb{P}\left[\hat{X}(Y) = 1 \mid X = 0\right]\Big)$$

Note that λ was arbitrary. So, we can maximize the RHS over all possible λ. Following the definition of $u(\beta)$, this implies

$$\underbrace{\mathbb{P}\left[\hat{X}(Y) = 0 \mid X = 1\right]}_{\text{type II error}} \geq u(\underbrace{\mathbb{P}\left[\hat{X}(Y) = 1 \mid X = 0\right]}_{\text{type I error}})$$

This implies that all $\hat{X}$ lie above or on the error curve. However, since all threshold tests lie on or below the error curve as well, combining these bounds indicates that all threshold tests lie exactly on the error curve. If you ignore all the clunky notation, the previous statement is basically saying that $y \leq u(x)$ and $y \geq u(x)$ both imply $y = u(x)$.

Where does randomization enter the picture? Threshold tests don't always continuously sweep out the curve u. For example,



The blue parts of the error curve above are not achieved by any threshold test. Then, how do you achieve something like the red point? We can just flip a biased (the randomization constant γ) coin to decide between the tests on either side of it.

21.2 Estimation

Hypothesis testing tries to discriminate between 2 (or more) discrete hypotheses. Estimation is another inference problem, but now we try to guess the numerical value of some unknown quantity. For example,

- Given measurements by GPS, what is my latitude/longitude?
- Given the history of a stock, what will be the value tomorrow?
- A lot of common problems in ML, communications, signal processing, finance, etc.

The basic setup for estimation goes as follows

$$\underbrace{X}_{\text{unknown RV}} \longrightarrow \underbrace{P_{Y|X}}_{\text{model}} \longrightarrow \underbrace{Y}_{\text{observation(s)}} \longrightarrow \text{estimation procedure} \longrightarrow \underbrace{\hat{X}(y)}_{\text{estimate}}$$

Like hypothesis testing, the model is assumed to be given. The goal of estimation problems is to choose an $\hat{X}$ that minimizes $\mathbb{E}\left[(X - \hat{X}(Y))^2\right]$, i.e., mean square error (MSE).

Definition 21.1: MMSE
The minimum mean squared estimator (MMSE) of X given Y is

$$\operatorname*{argmin}_{\hat{X}} \mathbb{E}\left[\left(X - \hat{X}(Y)\right)^2\right] = \mathbb{E}[X \mid Y]$$

Proof: Let $\phi(Y)$ be the MMSE of X given Y. Then,

$$\begin{aligned}\mathbb{E}\left[(X - \phi(Y))^2\right] &= \mathbb{E}\left[(X - \mathbb{E}[X \mid Y] + \mathbb{E}[X \mid Y] - \phi(Y))^2\right] \\ &= \mathbb{E}\left[(X - \mathbb{E}[X \mid Y])^2\right] + 2\mathbb{E}[(X - \mathbb{E}[X \mid Y])(\mathbb{E}[X \mid Y] - \phi(Y))] + \mathbb{E}\left[(\mathbb{E}[X \mid Y] - \phi(Y))^2\right]\end{aligned}$$

Let $f(Y) = \mathbb{E}[X \mid Y] - \phi(Y)$. Note that,

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X \mid Y]\,(\mathbb{E}[X \mid Y] - \phi(Y))] &= \mathbb{E}[\mathbb{E}[X \mid Y] f(Y)] \\ &= \sum_y \mathbb{E}[X \mid Y = y] f(y) \mathbb{P}[Y = y] \\ &= \sum_y \left(\sum_x x \mathbb{P}[X = x \mid Y = y]\right) f(y) \mathbb{P}[Y = y] \\ &= \sum_{x,y} x f(y) \mathbb{P}[X = x, Y = y] \\ &= \mathbb{E}[X f(Y)] \\ &= \mathbb{E}[X(\mathbb{E}[X \mid Y] - \phi(Y))]\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}[X \mid Y])(\mathbb{E}[X \mid Y] - \phi(Y))] &= \mathbb{E}[X(\mathbb{E}[X \mid Y] - \phi(Y))] - \mathbb{E}[\mathbb{E}[X \mid Y]\,(\mathbb{E}[X \mid Y] - \phi(Y))] \\ &= 0\end{aligned}$$

Then,

$$\mathbb{E}\left[(X - \phi(Y))^2\right] = \mathbb{E}\left[(X - \mathbb{E}[X \mid Y])^2\right] + \mathbb{E}\left[(\mathbb{E}[X \mid Y] - \phi(Y))^2\right]$$

Both of the terms on the right are non-zero so the LHS can only be minimized if one of them goes to 0. Note that the expression on the right is equivalent to the LHS when $\phi(Y) = \mathbb{E}[X \mid Y]$. Therefore, the MMSE is achieved when using the conditional expectation $\mathbb{E}[X \mid Y]$.

So, the MSE-based estimation problem is totally solved in this sense. However, it is not really practical in many cases

1. $\mathbb{E}[X \mid Y]$ can be intractable to compute, even if P_{XY} is known exactly (usually involves integration or can just be very high complexity)
2. Often, we don't know P_{XY} exactly but only have some reasonable model of it

Workaround: focus on linear estimation and minimize MSE over all "linear estimators" of the form

$$\hat{X}(Y) = a + \sum_{i=1}^{n} b_i Y_i$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is a vector of observations. This problem is called linear least squares (LLS) estimation. The best linear estimator is called the linear least squares estimator (LLSE), and is denoted by

$$\mathbb{L}[X \mid \mathbf{Y}]$$

since the notation reminds us of $\mathbb{E}[X \mid Y]$, the minimum mean squared estimator, but is for linear estimators. As we will see over the next few lectures, LLS estimation is just linear algebra disguised as probability.

22 Lecture 22

22.1 Linear Estimation

Definition 22.1: LLSE

$$\mathbb{L}[X \mid \mathbf{Y}] = \underset{\text{linear } \hat{X}}{\operatorname{argmin}} \mathbb{E}\left[\left|X - \hat{X}(Y)\right|^2\right]$$

where linear estimators $\hat{X}(\mathbf{Y})$ are of the form

$$\hat{X}(\mathbf{Y}) = a + \sum_{i=1}^{n} b_i Y_i$$

for $a, b_i \in \mathbb{R}$

How do you solve for a and $b_1, \ldots, b_n$ in the following?

$$\mathbb{L}[X \mid \mathbf{Y}] = \min_{a, b_1, \ldots, b_n} \mathbb{E}\left[\left|X - \left(a + \sum_i b_i Y_i\right)\right|^2\right]$$

Here is a calculus based approach:

$$\begin{aligned} J(a, b_1, \ldots, b_n) &= \mathbb{E}\left[\left|X - \left(a + \sum b_i Y_i\right)\right|^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2a\mathbb{E}[X] - 2\sum_i b_i \mathbb{E}[XY_i] + a^2 + 2a\sum_i b_i \mathbb{E}[Y_i] + \sum_i b_i^2 \mathbb{E}\left[Y_i^2\right] + \sum_{i \neq j} b_i b_j \mathbb{E}\left[Y_i Y_j\right] \end{aligned}$$

Then,

$$\begin{aligned} \frac{\partial J}{\partial a} = 0 &\implies a = \mathbb{E}[X] - \sum_{i=1}^{n} b_i \mathbb{E}[Y_i] \\ \frac{\partial J}{\partial b_i} = 0 &\implies \mathbb{E}[XY_i] = a\mathbb{E}[Y_i] + b_i \mathbb{E}\left[Y_i^2\right] + \sum_{j \neq i} b_j \mathbb{E}\left[Y_i Y_j\right] \end{aligned}$$

This is a system of linear equations that we can solve for $a, b_1, \ldots, b_n$. Let's make life easy by assuming $\mathbb{E}[X] = \mathbb{E}[Y_i] = 0$ for all $i = 1, \ldots n$. Then,

$$\begin{aligned} a &= 0 \\ \mathbb{E}[XY_i] &= \sum_{j=1}^{n} b_j \mathbb{E}\left[Y_i Y_j\right], \quad i \in \{1, \ldots, n\} \end{aligned}$$

Definition 22.2: Covariance and Cross-Covariance Matrices

Define

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{XY}} &= \mathbb{E}\left[(X - \mu_X)(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})^T\right] \quad \text{(cross-covariance matrix)} \\ \boldsymbol{\Sigma}_{\mathbf{Y}} &= \mathbb{E}\left[(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})^T\right] \quad \text{(covariance matrix)} \end{aligned}$$

where $\mu_X = \mathbb{E}[X]$ and $\boldsymbol{\mu}_{\mathbf{Y}} = \mathbb{E}[\mathbf{Y}] = \begin{bmatrix}\mathbb{E}[Y_1] & \ldots & \mathbb{E}[Y_n]\end{bmatrix}^T$.

Thus,

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{XY}} &= \mathbf{b}^T \boldsymbol{\Sigma}_{\mathbf{Y}} \\ \mathbf{b}^T &= \boldsymbol{\Sigma}_{\mathbf{XY}} \boldsymbol{\Sigma}_{\mathbf{Y}}^{-1} \end{aligned}$$

In particular, when $\mathbb{E}[X] = \mathbb{E}[Y_i] = 0$, then $\mathbb{L}[X \mid \mathbf{Y}] = \mathbf{b}^T \mathbf{Y} = \boldsymbol{\Sigma}_{\mathbf{XY}} \boldsymbol{\Sigma}_{\mathbf{Y}}^{-1} \mathbf{Y}$. If they are not zero mean, just add the means back in to get

$$\mathbb{L}[X \mid \mathbf{Y}] = \mu_X + \boldsymbol{\Sigma}_{\mathbf{XY}} \boldsymbol{\Sigma}_{\mathbf{Y}}^{-1} (\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})$$

Note 22.1

If $\mathbf{X} = \begin{bmatrix} X_1 & \dots & X_n \end{bmatrix}^T$ is a vector, then the linear estimation problem is of the form:

$$\min_{\text{linear } \hat{\mathbf{X}}} \mathbb{E}\left[\left\|\mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y})\right\|_2^2\right] = \sum_{i=1}^{k} \min_{\text{linear } \hat{X}_i} \mathbb{E}\left[\left|X_i - \hat{X}_i(\mathbf{Y})\right|^2\right]$$

So, we can always assume that X is a scalar because the vector problems just decompose into scalar problems for each vector component. Nevertheless, the expression above remains valid for vector-valued X as well.

Observe that $\mathbb{L}[X \mid Y]$ only depends on the first and second order statistics of X and Y (means and covariances). In practice, this is good because we rarely know the joint distribution of X, Y completely but we can estimate the first/second-order statistics from the data.

Moreover, the *linear* in linear estimation comes from the fact that the estimator is linear in $a, b_1, \dots, b_n$ but not necessarily X or Y.

Example 22.1: QLSE

Let X, Y be random variables. Compute the best quadratic estimator (QLSE) of the form

$$\hat{X}(Y) = a + b_1 Y + b_2 Y^2, \forall a, b_1, b_2 \in \mathbb{R}$$

Since the QLSE is still linear in a, b_1, b_2, it is just

$$\mathbb{L}\left[X \mid \tilde{\mathbf{Y}}\right] = \mu_X + \boldsymbol{\Sigma}_{\mathbf{X}\tilde{\mathbf{Y}}} \boldsymbol{\Sigma}_{\tilde{\mathbf{Y}}}^{-1} (\tilde{\mathbf{Y}} - \boldsymbol{\mu}_{\tilde{\mathbf{Y}}})$$

where $\tilde{\mathbf{Y}} = \begin{bmatrix} Y \\ Y^2 \end{bmatrix}$.

22.2 Connection to Linear Regression

Consider a sample observation model

$$\begin{aligned} \mathbf{Y} &= \mathbf{A}\mathbf{X} + \mathbf{Z} \\ \boldsymbol{\Sigma}_{\mathbf{X}} &= \sigma_X^2 \mathbf{I} \\ \boldsymbol{\Sigma}_{\mathbf{Z}} &= \sigma_Z^2 \mathbf{I} \\ \mathbf{A} &\in \mathbb{R}^{n \times k} \end{aligned}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are uncorrelated, $\mathbf{Y}$ is an n-vector and $\mathbf{X}$ is a k-vector. Assume that everything is zero-mean for simplicity. Then, the best linear estimator of $\mathbf{X}$ given $\mathbf{Y}$ is

$$\mathbb{L}[\mathbf{X} \mid \mathbf{Y}] = \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{Y}} \boldsymbol{\Sigma}_{\mathbf{Y}}^{-1} \mathbf{Y}$$

Note that

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{Y}} &= \mathbb{E}\left[\mathbf{X}\mathbf{Y}^T\right] \\ &= \mathbb{E}\left[\mathbf{X}\left(\mathbf{X}^T\mathbf{A}^T + \mathbf{Z}^T\right)\right] \\ &= \sigma_X^2 \mathbf{A}^T \\ \boldsymbol{\Sigma}_{\mathbf{Y}} &= \mathbb{E}\left[\mathbf{Y}\mathbf{Y}^T\right] \\ &= \mathbb{E}\left[(\mathbf{A}\mathbf{X} + \mathbf{Z})(\mathbf{A}\mathbf{X} + \mathbf{Z})^T\right] \\ &= \sigma_X^2 \mathbf{A}\mathbf{A}^T + \sigma_Z^2 \mathbf{I} \\ \mathbb{L}[\mathbf{X} \mid \mathbf{Y}] &= \sigma_X^2 \mathbf{A}^T \left(\sigma_X^2 \mathbf{A}\mathbf{A}^T + \sigma_Z^2 \mathbf{I}\right)^{-1} \mathbf{Y} \end{aligned}$$

$$= \mathbf{A}^T \left(\mathbf{A}\mathbf{A}^T + \frac{\sigma_Z^2}{\sigma_X^2} \mathbf{I} \right)^{-1} \mathbf{Y}$$

This is the result of a "Bayesian setting" where I assume I know σ_X^2. If I don't know σ_X^2, then the best I can do is assume that $\sigma_X^2 = \infty$. This, implies that

$$\mathbb{L}[\mathbf{X} \mid \mathbf{Y}] = \mathbf{A}^T \left(\mathbf{A}\mathbf{A}^T \right)^{-1} \mathbf{Y}$$

In general,

$$\mathbb{L}[\mathbf{X} \mid \mathbf{Y}] = \mathbf{A}^\dagger \mathbf{Y}$$

where $\mathbf{A}^\dagger$ is the pseudoinverse of matrix $\mathbf{A}$. For a full row rank matrix,

$$\mathbf{A}^\dagger = \lim_{\gamma \to +\infty} \mathbf{A}^T \left(\gamma^{-1} \mathbf{I} + \mathbf{A}\mathbf{A}^T \right)^{-1}$$

When $\mathbf{A}$ is full column rank instead (this is the overdetermined case which is of the most interest in least-squares regression problems anyways), then $\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ giving us $\mathbb{L}[\mathbf{X} \mid \mathbf{Y}] = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}$.

Recall that the linear regression problem $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$ has the solution given by $\hat{\mathbf{x}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^{-1}\mathbf{y}$ (from EECS 16AB). Moral of the story is that linear regression can be considered a special case of linear estimation (with a non-Bayesian, linear observation model).

22.3 Geometry of Linear Estimation

Let $\mathcal{V}$ be a vector space over a real scalar field. Let $\langle \cdot, \cdot \rangle$ be an inner product on $\mathcal{V}$. That is,

1. $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$ for $\mathbf{u}, \mathbf{v} \in \mathcal{V}$
2. $\langle a\mathbf{u} + b\mathbf{v}, \mathbf{w} \rangle = a \langle \mathbf{u}, \mathbf{w} \rangle + b \langle \mathbf{v}, \mathbf{w} \rangle$ for $a, b \in \mathbb{R}$, $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathcal{V}$
3. $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$, $\forall \mathbf{u} \in \mathcal{V}$ and $\langle \mathbf{u}, \mathbf{u} \rangle = 0 \iff \mathbf{u} = \mathbf{0}$

The vector space $\mathcal{V}$ when paired with an inner product $\langle \cdot, \cdot \rangle$ is also called a (real) inner product space. It is a normed vector space, with norm $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$. Norms satisfy the following properties

- $\|a\mathbf{v}\| = |a|\|\mathbf{v}\|$ for $a \in \mathbb{R}$, $\mathbf{v} \in \mathcal{V}$
- $\|\mathbf{v}\| \geq 0$ and $\|\mathbf{v}\| = 0 \iff \mathbf{v} = \mathbf{0}$
- $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$

Then the inner product space $\mathcal{V}$ is called a "Hilbert Space" if it is "complete" with respect to the norm $\|\cdot\|$ (completeness means we can take limits without popping out of the space).

Example 22.2

The space of continuous bounded functions on $\mathbb{R}$, denoted by $C_b(\mathbb{R})$, is complete with respect to the norm $\|f\| = \max_x |f(x)|$ but is not complete with respect to $\|f\| = \int |f|^2 \, \mathrm{d}x$.

Hilbert spaces enjoy a notion of geometry compatible with our intuition.

Theorem 22.1: Hilbert Projection Theorem

Let $\mathcal{H}$ be a Hilbert Space, and $\mathcal{U} \subseteq \mathcal{H}$ be a closed subspace. For each $\mathbf{v} \in \mathcal{H}$, there is a unique closest point $\mathbf{u} \in \mathcal{U}$ to $\mathbf{v}$, i.e.,

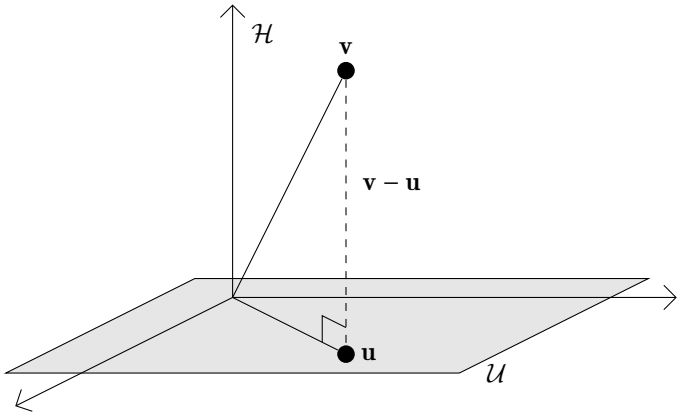
$$\operatorname*{argmin}_{\mathbf{u} \in \mathcal{U}} \|\mathbf{u} - \mathbf{v}\|$$

exists and is unique. Moreover, $\mathbf{u} \in \mathcal{U}$ is the closest point to $\mathbf{v} \in \mathcal{H}$ if and only if $\langle \mathbf{u} - \mathbf{v}, \mathbf{u}' \rangle = 0$, $\forall \mathbf{u}' \in \mathcal{U}$ (i.e. $\mathbf{u} - \mathbf{v}$ and $\mathbf{u}'$ are orthogonal).

Note that we can derive the Pythagorean theorem using inner products

Proof:

$$\begin{aligned}\|\mathbf{u}\|^2 + \|\mathbf{u}-\mathbf{v}\|^2 &= \langle \mathbf{u},\mathbf{u}\rangle + \langle \mathbf{u}-\mathbf{v},\mathbf{u}-\mathbf{v}\rangle \\ &= \langle \mathbf{u},\mathbf{u}\rangle + \langle \mathbf{u}-\mathbf{v},\mathbf{u}\rangle - \langle \mathbf{u}-\mathbf{v},\mathbf{v}\rangle \\ &= \langle \mathbf{u}-\mathbf{v},\mathbf{u}\rangle + (\langle \mathbf{u},\mathbf{u}\rangle - \langle \mathbf{u},\mathbf{v}\rangle) + \langle \mathbf{v},\mathbf{v}\rangle \\ &= 2\langle \mathbf{u}-\mathbf{v},\mathbf{u}\rangle + \langle \mathbf{v},\mathbf{v}\rangle \\ &= \langle \mathbf{v},\mathbf{v}\rangle \\ &= \|\mathbf{v}\|^2\end{aligned}$$



A pictorial illustration of the Hilbert Projection Theorem

22.4 Orthogonality Principle

Theorem 22.2: Hilbert Space of Random Variables
Let $(\Omega, \mathcal{F}, P)$ be a probability space. The collection of random variables X with $\mathbb{E}\left[X^2\right] < \infty$ (i.e. finite second moments) form a Hilbert space with respect to the inner product $\langle X, Y\rangle = \mathbb{E}[XY]$.

Note 22.2
In this notation, $\|X\|^2 = \langle X, X\rangle = \mathbb{E}\left[X^2\right]$.

For random variables $Y_1, Y_2, \ldots, Y_n$ with finite second moments, the space of RVs

$$\mathcal{U} = \left\{a + \sum b_i Y_i \mid a, b_i \in \mathbb{R}\right\}$$

is a closed subspace of the Hilbert space of all RVs $\mathcal{H}$. By the Hilbert Projection Theorem,

$$\begin{aligned}\mathbb{L}[X \mid \mathbf{Y}] &= \operatorname*{argmin}_{u\in\mathcal{U}} \|X-u\|^2 \\ &= \operatorname*{argmin}_{\text{linear } \hat{X}} \mathbb{E}\left[\left|X - \hat{X}(\mathbf{Y})\right|^2\right]\end{aligned}$$

exists and is unique. Moreover, it is characterized by the following equations:

$$\begin{aligned}\langle \mathbb{L}[X \mid \mathbf{Y}] - X, u\rangle &= 0 \\ \mathbb{E}[(\mathbb{L}[X \mid \mathbf{Y}] - X)u] &= 0, \forall u \in \mathcal{U} \\ \mathbb{E}\left[(\mathbb{L}[X \mid \mathbf{Y}] - X)\left(a + \sum b_i Y_i\right)\right] &= 0, \forall a, b_1, \ldots, b_n \in \mathbb{R}\end{aligned}$$

$$\mathbb{E}[\mathbb{L}[X \mid \mathbf{Y}]] = \mathbb{E}[X]$$
$$\mathbb{E}[\mathbb{L}[X \mid \mathbf{Y}]Y_i] = \mathbb{E}[XY_i],\ i = 1, \ldots, n$$

The last two equations form the "basis" of $\mathcal{U}$ (by selectively choosing values of $a, b_1, \ldots, b_n$ as 0 or 1). This is called the "orthogonality principle" since it uniquely characterizes $\mathbb{L}[X \mid \mathbf{Y}]$. To see this agrees with what we derived before, try plugging in

$$\mathbb{L}[X \mid \mathbf{Y}] = \mu_X + \mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})$$

to get

$$\begin{aligned}
\mathbb{E}[\mathbb{L}[X \mid \mathbf{Y}]] &= \mathbb{E}\Big[\mu_X + \mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})\Big] \\
&= \mu_X + \mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1}\mathbb{E}[\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y}] \\
&= \mu_X \\
&= \mathbb{E}[X]
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{L}[X \mid \mathbf{Y}]\mathbf{Y}^T\Big] &= \mathbb{E}\Big[\mathbb{L}[X \mid \mathbf{Y}]\,(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})^T\Big] + \mathbb{E}\Big[\mathbb{L}[X \mid \mathbf{Y}]\boldsymbol{\mu}_\mathbf{Y}{}^T\Big] \\
&= \mathbb{E}\Big[\Big(\mu_X + \mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})\Big)(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})^T\Big] + \mathbb{E}[\mathbb{L}[X \mid \mathbf{Y}]]\boldsymbol{\mu}_\mathbf{Y}{}^T \\
&= \mathbb{E}\Big[\mu_X(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})^T\Big] + \mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1}\mathbb{E}\Big[(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})^T\Big] + \mu_X\boldsymbol{\mu}_\mathbf{Y}{}^T \\
&= \mathbb{E}\Big[\mu_X\mathbf{Y}^T\Big] - \mu_X\boldsymbol{\mu}_\mathbf{Y}{}^T + \mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1}\mathbf{\Sigma_Y} + \mu_X\boldsymbol{\mu}_\mathbf{Y}{}^T \\
&= \mathbb{E}\Big[\mu_X\mathbf{Y}^T\Big] + \mathbf{\Sigma_{XY}} \\
&= \mathbb{E}\Big[\mu_X\mathbf{Y}^T\Big] + \mathbb{E}\Big[(X - \mu_X)\mathbf{Y}^T\Big] \\
&= \mathbb{E}\Big[X\mathbf{Y}^T\Big]
\end{aligned}$$

23 Lecture 23

23.1 Orthogonality Principle Cont.

The orthogonality principle gives us a characterization of $\mathbb{L}[X \mid \mathbf{Y}]$ as

$$\mathbb{E}[\mathbb{L}[X \mid \mathbf{Y}]] = \mathbb{E}[X]$$
$$\mathbb{E}[(\mathbb{L}[X \mid \mathbf{Y}] - X)Y_i] = 0, \quad i = 1, \ldots, n$$

The first equation means that LLSE is unbiased. The second equation indicates that LLSE error is orthogonal to all observations, i.e., error is uncorrelated with the observations. Following the Hilbert Projection Theorem, we also have

$$\mathbb{E}[(\mathbb{L}[X \mid \mathbf{Y}] - X)u] = 0, \forall u \in \mathcal{U}$$

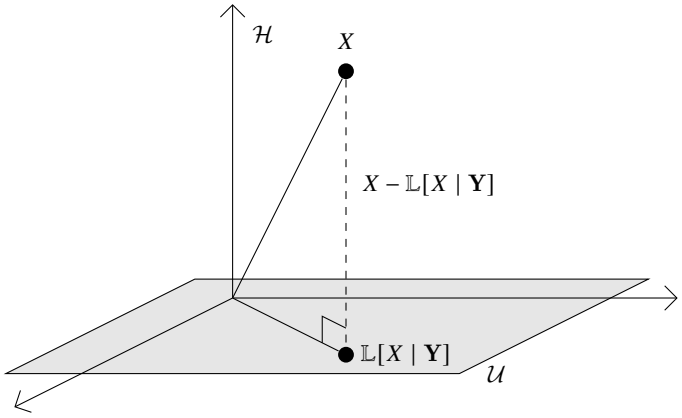
but each $u \in \mathcal{U}$ can be written as $a + \sum b_i Y_i$. Thus,

$$a\mathbb{E}[(\mathbb{L}[X \mid \mathbf{Y}] - X)] + \sum b_i \mathbb{E}[(\mathbb{L}[X \mid \mathbf{Y}] - X)Y_i] = 0, \forall a, b_i \in \mathbb{R}$$

Setting $\mathbb{L}[X \mid Y] = a + \sum b_i Y_i$, plugging into the orthogonality principle and solving gives us

$$\mathbb{L}[X \mid \mathbf{Y}] = \mu_X + \mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})$$

23.2 LLSE Error



The squared error of an LLSE estimator is

$$\begin{aligned}\mathbb{E}\Big[|X - \mathbb{L}[X \mid \mathbf{Y}]|^2\Big] &= \|X - \mathbb{L}[X \mid \mathbf{Y}]\|^2 \\ &= \|X\|^2 - \|\mathbb{L}[X \mid \mathbf{Y}]\|^2 \\ &= \mathbb{E}\Big[X^2\Big] - \mathbb{E}\Big[\mathbb{L}[X \mid \mathbf{Y}]^2\Big]\end{aligned}$$

The second equation follows from the Pythagorean theorem since $X - \mathbb{L}[X \mid \mathbf{Y}]$, $\mathbb{L}[X \mid \mathbf{Y}]$ and X form a right triangle with the former two as the legs (see the figure above). Since $\mathbb{L}[X \mid \mathbf{Y}]$ is unbiased, we can assume WLOG that $\mathbb{E}[X] = 0$. Then,

$$\begin{aligned}\mathbb{E}\Big[|X - \mathbb{L}[X \mid \mathbf{Y}]|^2\Big] &= \mathrm{Var}(X) - \mathbb{E}\Big[\Big|\mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})\Big|^2\Big] \\ &= \mathrm{Var}(X) - \mathbb{E}\Big[\mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})^T\mathbf{\Sigma_Y}^{-1}\mathbf{\Sigma_{XY}}^T\Big] \\ &= \mathrm{Var}(X) - \mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1}\underbrace{\mathbb{E}\Big[(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})^T\Big]}_{\mathbf{\Sigma_Y}}\mathbf{\Sigma_Y}^{-1}\mathbf{\Sigma_{XY}}^T\end{aligned}$$

$$= \text{Var}(X) - \mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1}\mathbf{\Sigma_{XY}}^T$$
$$= \text{Var}(X) - \mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1}\mathbf{\Sigma_{YX}}$$

Therefore, the entire theory of Linear Least Squared Estimation in a nutshell can be summarized by

$$\mathbb{L}[X \mid \mathbf{Y}] = \mu_X + \mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1}(\mathbf{Y} - \boldsymbol{\mu_Y})$$
$$\text{LLSE error} = \text{Var}(X) - \mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1}\mathbf{\Sigma_{YX}}$$

From here, it's just plug and chug into applications.

23.3 Applications of the Orthogonality Principle (MMSE)

Let X, Y be RVs and $\mathbb{E}\left[X^2\right] < \infty$. As shown before,

$$\mathbb{E}[X \mid Y] = \underset{\hat{X}(Y)}{\text{argmin}}\, \mathbb{E}\left[\left|X - \hat{X}(Y)\right|^2\right]$$

where $\hat{X}(Y)$ is any function of Y (not necessarily linear).
Recall that the "real" definition of conditional expectation, at least when we first encountered it, was the tower property given by

$$\mathbb{E}[\mathbb{E}[X \mid Y]g(Y)] = \mathbb{E}[Xg(Y)]$$

for all functions g of Y (say, with $\mathbb{E}\left[g(Y)^2\right] < \infty$). The orthogonality principle characterization of $\mathbb{E}[X \mid Y]$, defined in a similar way as LLS estimators using the Hilbert Projection Theorem, is precisely

$$\mathbb{E}[(X - \mathbb{E}[X \mid Y])g(Y)] = 0$$
$$\mathbb{E}[g(Y)\mathbb{E}[X \mid Y]] = \mathbb{E}[g(Y)X]$$

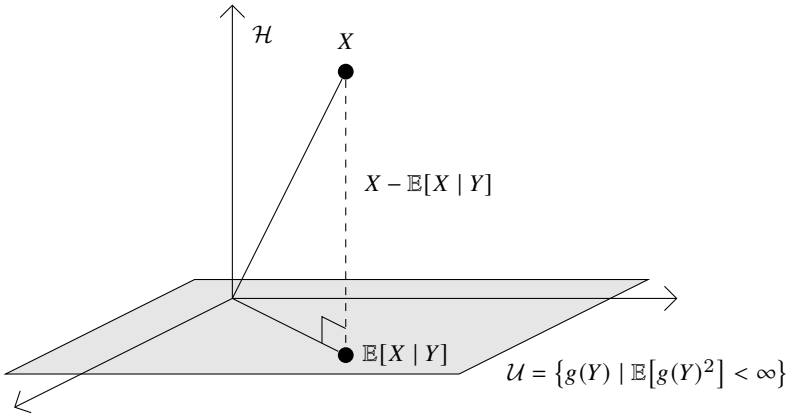
In other words, the conditional expectation is the projection of X onto the subspace of functions of Y. Thus, the tower property is the same as characterization via the orthogonality principle.

23.4 MMSE Error

What is the MMSE error? Using the same WLOG assumptions of zero mean and the Pythagorean theorem again,

$$\|X\|^2 = \|\mathbb{E}[X \mid Y]\|^2 + \|X - \mathbb{E}[X \mid Y]\|^2$$
$$\mathbb{E}\left[|X|^2\right] = \mathbb{E}\left[|\mathbb{E}[X \mid Y]|^2\right] + \mathbb{E}\left[|X - \mathbb{E}[X \mid Y]|^2\right]$$
$$\mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \left(\mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right] - \mathbb{E}[X]^2\right) + \mathbb{E}\left[(X - \mathbb{E}[X \mid Y])^2\right]$$
$$= \left(\mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right] - \mathbb{E}[\mathbb{E}[X \mid Y]]^2\right) + \mathbb{E}\left[\mathbb{E}\left[(X - \mathbb{E}[X \mid Y])^2 \mid X\right]\right]$$
$$\text{Var}(X) = \text{Var}(\mathbb{E}[X \mid Y]) + \mathbb{E}[\text{Var}(X \mid Y)]$$

Observe that the expression above is just the Law of Total Variance! Thus, the MMSE error is $\text{Var}(X) - \text{Var}(\mathbb{E}[X \mid Y])$.



Observe that in the picture above, $\mathcal{U}$ is the subspace of *all* functions of Y (such that they have a finite variance) instead of just the linear functions as we have been denoting it so far.

Note 23.1
In general, $\mathbb{L}[X \mid Y] \neq \mathbb{E}[X \mid Y]$. However, there are special cases where $\mathbb{L}[X \mid Y] = \mathbb{E}[X \mid Y]$, most notably when X, Y are jointly gaussian (more on this later) or when $\mathbb{E}[X \mid Y]$ happens to be linear with respect to Y.

23.5 Online estimation

In the real world, data is observed sequentially. How do we efficiently update our LLS estimate on the arrival of new observations?

For motivation, start with a simple setting. WLOG, assume $\mathbb{E}[X] = 0$. Suppose observations $Y_1, Y_2, \ldots$ are orthogonal, i.e., $\langle Y_i, Y_j \rangle = 0$ for $i \neq j$ to each other. For convenience, define

$$\begin{aligned} \mathbf{Y}^n &= (Y_1, \ldots, Y_n) \\ \mu_i &= \mathbb{E}[Y_i] \end{aligned}$$

Then,

$$\begin{aligned} \mathbb{L}\left[X \mid \mathbf{Y}^{n+1}\right] &= \mathbb{L}\left[X \mid \mathbf{Y}^n\right] + \mathbb{L}[X \mid Y_{n+1}] \\ &= \mathbb{L}\left[X \mid \mathbf{Y}^n\right] + \frac{\mathrm{Cov}(X, Y_{n+1})}{\mathrm{Var}(Y_{n+1})}(Y_{n+1} - \mu_{n+1}) \end{aligned}$$

Note 23.2
Orthogonal means $\langle Y_i, Y_j \rangle = \mathbb{E}[Y_i Y_j] = 0$ and uncorrelated means $\mathrm{Cov}(Y_i, Y_j) = 0$. These coincide if the Y_is have zero mean, which we can always assume by simply centering around the observations.

Proof: Check the orthogonality principle!

$$\begin{aligned} \mathbb{E}\left[\left(\mathbb{L}\left[X \mid \mathbf{Y}^{n+1}\right] - X\right)Y_k\right] &= 0, \quad k = 1, \ldots, n+1 \\ \mathbb{E}\left[\left(\mathbb{L}\left[X \mid \mathbf{Y}^{n+1}\right] - X\right)Y_k\right] &= \mathbb{E}[(\mathbb{L}[X \mid \mathbf{Y}^n] + \mathbb{L}[X \mid Y_{n+1}] - X)Y_k] \end{aligned}$$

- If $k = n+1$,

$$\begin{aligned} \mathbb{E}[(\mathbb{L}[X \mid \mathbf{Y}^n] + \mathbb{L}[X \mid Y_{n+1}] - X)Y_k] &= \mathbb{E}[(\mathbb{L}[X \mid \mathbf{Y}^n] + \mathbb{L}[X \mid Y_{n+1}] - X)Y_{n+1}] \\ &= \underbrace{\mathbb{E}[(\mathbb{L}[X \mid Y_{n+1}] - X)Y_{n+1}]}_{=0 \text{ by orthgonality principle}} + \underbrace{\mathbb{E}[\mathbb{L}[X \mid \mathbf{Y}^n]Y_{n+1}]}_{=0 \text{ since } \langle Y_i, Y_{n+1} \rangle = 0} \end{aligned}$$

  The above follows since $\mathbb{L}[X \mid \mathbf{Y}^n]$ is a linear function of $Y_1, \ldots, Y_n$, and hence, must be orthogonal to Y_{n+1} too.

- If $1 \leq k \leq n$,

$$\mathbb{E}[(\mathbb{L}[X \mid \mathbf{Y}^n] + \mathbb{L}[X \mid Y_{n+1}] - X)Y_k] = \underbrace{\mathbb{E}[(\mathbb{L}[X \mid \mathbf{Y}^n] - X)Y_k]}_{=0 \text{ by orthgonality principle}} + \underbrace{\mathbb{E}[\mathbb{L}[X \mid Y_{n+1}]Y_k]}_{=0 \text{ since } \langle Y_{n+1}, Y_k \rangle = 0}$$

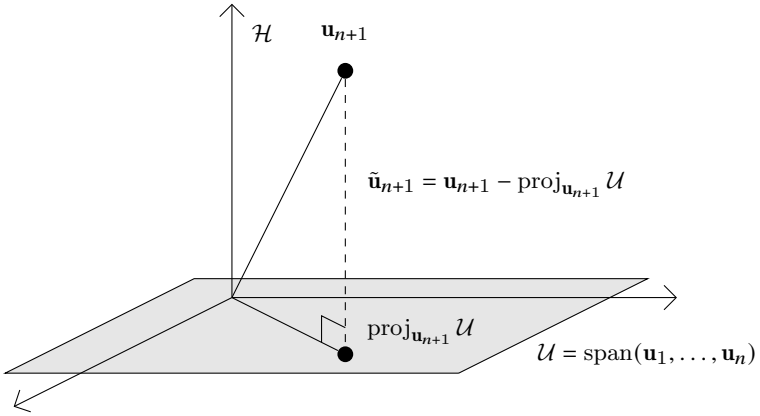
Main idea: showing that the equation satisfies the orthogonality principle lets us avoid tedious computations with matrices.

24 Lecture 24

24.1 Gram Schmidt for Random Variables

Last time, we came up with a procedure for online estimation. If the observations are uncorrelated (orthogonal), we have a ridiculously nice way of sequentially updating our estimate of X given any new observations.

In reality, however, observations are not uncorrelated (not orthogonal): if they are correlated, we can transform them to be uncorrelated using Gram-Schmidt. In linear algebra, we can take non-orthogonal vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ and make them orthogonal by running the Gram-Schmidt process. The resulting sequence $\tilde{\mathbf{u}}_1, \tilde{\mathbf{u}}_2, \ldots, \tilde{\mathbf{u}}_n$ is orthogonal and $\text{span}(\tilde{\mathbf{u}}_1, \tilde{\mathbf{u}}_2, \ldots, \tilde{\mathbf{u}}_n) = \text{span}(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$ for all $n \geq 1$.



The Gram Schmidt process for random variables goes as follows: given a sequence of RVs $Y_1, \ldots, Y_n$, define the update step

$$\tilde{Y}_{n+1} = Y_{n+1} - \mathbb{L}\left[Y_{n+1} \mid \mathbf{Y}^n\right]$$

where the final term is the projection of Y_{n+1} onto $\text{span}(1, Y_1, \ldots, Y_n)$. The resulting $\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_n$ are uncorrelated by the Gram-Schmidt construction (orthogonality principle) and

$$\text{span}(1, \tilde{Y}_1, \ldots, \tilde{Y}_n) = \text{span}(1, Y_1, \ldots, Y_n)$$

So, this means that

$$\mathbb{L}\left[X \mid \mathbf{Y}^n\right] = \mathbb{L}\left[X \mid \tilde{\mathbf{Y}}^n\right], \forall n \geq 1$$

The key thing gained here is that the $\tilde{Y}_i$s are uncorrelated (orthogonal) so we can sequentially compute $\mathbb{L}\left[X \mid \tilde{\mathbf{Y}}^n\right]$, and therefore, $\mathbb{L}[X \mid \mathbf{Y}^n]$.

Definition 24.1: Linear Innovation Sequence
The sequence $\tilde{Y}_1, \tilde{Y}_2, \ldots$ is called the linear innovation sequence (orthogonal innovations) corresponding to $Y_1, Y_2, \ldots$.

Here, $\tilde{Y}_n$ is the component of Y_n that can't be linearly estimated from $Y_1, \ldots, Y_n$ (equivalently $\tilde{Y}_1, \ldots, \tilde{Y}_n$). By transforming

$$Y_1, Y_2, \cdots \longrightarrow \tilde{Y}_1, \tilde{Y}_2, \ldots$$

we can always do sequential updates to our linear estimator using the simple process we saw earlier.

Note 24.1
For any three zero-mean RVs X, Y, Z, where Y and Z are uncorrelated, the following holds

$$\mathbb{L}[X \mid Y, Z] = \mathbb{L}[X \mid Y] + \mathbb{L}[X \mid Z]$$

The notation $\mathbb{L}[X \mid Y, Z]$ is describing the best linear estimator for X given both observations Y and Z. However, if Y and

Z are not uncorrelated, we can use an orthogonal innovation to define $\tilde{Z} = \mathbb{L}[Z \mid Y]$. Now Y and $\tilde{Z}$ are uncorrelated and span the same space of RVs as Y and Z. Then,

$$\begin{aligned}\mathbb{L}[X \mid Y, Z] &= \mathbb{L}[X \mid Y] + \mathbb{L}\left[X \mid \tilde{Z}\right] \\ &= \mathbb{L}[X \mid Y] + \mathbb{L}[X \mid Z - \mathbb{L}[Z \mid Y]]\end{aligned}$$

24.2 Jointly Gaussian Random Variables

Definition 24.2: Gaussian Vector

A gaussian random vector $\mathbf{X} = (X_1, \ldots, X_n)^T$ with density on $\mathbb{R}^n$ is defined via its PDF

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma}_{\mathbf{X}})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})^T \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})}$$

We write $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}})$ for short where $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^T\right]$ is the covariance matrix and $\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}[\mathbf{X}]$ is the mean vector (element wise mean of each component).

Note 24.2

Gaussian vectors have distributions parameterized by only the mean and covariance, just like the 1D case.

There are many equivalent definitions of gaussian random vectors, including

- Definition via the PDF as above
- Gaussian random vectors are affine transformations of IID gaussian random variables. In other words, if $\mathbf{X}$ has nonsingular $\boldsymbol{\Sigma}_{\mathbf{X}}$, then we can write

  $$\mathbf{X} = \boldsymbol{\mu}_{\mathbf{X}} + \mathbf{A}\mathbf{W}$$

  for some full rank $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{W} = (W_1, \ldots, W_n)^T$, where $W_i \sim_{\text{IID}} \mathcal{N}(0, 1)$. Why? Using derived distributions,

  $$\begin{aligned}f_{\mathbf{X}}(\mathbf{x}) &= \frac{1}{|\det(\mathbf{A})|} f_{\mathbf{W}}(\mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})) \\ &= \frac{1}{|\det(\mathbf{A})|} \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}|\mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})|^2} \\ &= \frac{1}{\left|\det\left(\mathbf{A}\mathbf{A}^T\right)\right|^{\frac{1}{2}}} \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})^T(\mathbf{A}\mathbf{A}^T)^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})}\end{aligned}$$

  But

  $$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{X}} &= \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^T\right] \\ &= \mathbb{E}\left[\mathbf{A}\mathbf{W}\mathbf{W}^T\mathbf{A}^T\right] \\ &= \mathbf{A}\mathbf{I}\mathbf{A}^T \\ &= \mathbf{A}\mathbf{A}^T\end{aligned}$$

- $\mathbf{X}$ is a gaussian random vector iff all one-dimensional projections of it are gaussian random variables. That is,

  $$\mathbf{a}^T\mathbf{X} \sim \mathcal{N}(\mathbf{a}^T\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{a}), \forall \mathbf{a} \in \mathbb{R}^n$$

For random vectors $\mathbf{X}, \mathbf{Y}$, we can partition the covariance matrix of their joint distribution (formed by stacking them) as

$$\text{Cov}\left(\begin{bmatrix}\mathbf{X} \\ \mathbf{Y}\end{bmatrix}\right) = \begin{bmatrix}\mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^T\right] & \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})^T\right] \\ \mathbb{E}\left[(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^T\right] & \mathbb{E}\left[(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})^T\right]\end{bmatrix}$$

$$= \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}$$

Theorem 24.1: Most Amazing Fact

If $\mathbf{X}$ and $\mathbf{Y}$ are jointly gaussian vectors, then we can always decompose $\mathbf{X}$ as

$$\mathbf{X} = \mu_X + \Sigma_{XY}{\Sigma_Y}^{-1}(\mathbf{Y} - \mu_Y) + V$$

where $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \Sigma_X - \Sigma_{XY}{\Sigma_Y}^{-1}\Sigma_{YX})$ is independent of $\mathbf{Y}$. This can also be applied to $\mathbf{Y}$ or any subset of components in a random vector.

Proof: Let $\tilde{\mathbf{X}} = \mu_X + \Sigma_{XY}{\Sigma_Y}^{-1}(\mathbf{Y} - \mu_Y) + \mathbf{V}$. Show that $\tilde{\mathbf{X}}$ and $\mathbf{Y}$ are jointly gaussian. Since $\mathbf{Y}$ and $\mathbf{V}$ are independent gaussians, they can be rewritten as

$$\mathbf{Y} = \mu_Y + \mathbf{A}\mathbf{W}_1$$
$$\mathbf{V} = \mathbf{B}\mathbf{W}_2$$

for some $\mathbf{A}, \mathbf{B}$ and $\mathbf{W}_1, \mathbf{W}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ (with appropriately sized $\mathbf{I}$ for each vector) independent of each other. This implies that

$$\begin{bmatrix} \tilde{\mathbf{X}} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} + \begin{bmatrix} \Sigma_{XY}{\Sigma_Y}^{-1}\mathbf{A} & \mathbf{B} \\ \mathbf{A} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix}$$

which is an affine transformation of IID standard gaussians. Since $(\tilde{\mathbf{X}}, \mathbf{Y})$ is jointly gaussian, its distribution is parameterized entirely by its mean and covariance,

$$\begin{aligned}
\mathbb{E}[\tilde{\mathbf{X}}] &= \mu_X \\
\mathbb{E}[\mathbf{Y}] &= \mu_Y \\
\Sigma_{\tilde{X}Y} &= \mathbb{E}\left[\left(\tilde{\mathbf{X}} - \mu_X\right)(\mathbf{Y} - \mu_Y)^T\right] \\
&= \mathbb{E}\left[\left(\Sigma_{XY}{\Sigma_Y}^{-1}(\mathbf{Y} - \mu_Y) + \mathbf{V}\right)(\mathbf{Y} - \mu_Y)^T\right] \\
&= \mathbb{E}\left[\Sigma_{XY}\Sigma_Y(\mathbf{Y} - \mu_Y)(\mathbf{Y} - \mu_Y)^{-1}\right] + \mathbb{E}\left[\mathbf{V}(\mathbf{Y} - \mu_Y)^T\right] \\
&= \Sigma_{XY}{\Sigma_Y}^{-1}\mathbb{E}\left[(\mathbf{Y} - \mu_Y)(\mathbf{Y} - \mu_Y)^T\right] + \mathbb{E}[\mathbf{V}]\mathbb{E}\left[(\mathbf{Y} - \mu_Y)^T\right] \\
&= \Sigma_{XY} \\
\Sigma_{\tilde{X}} &= \mathbb{E}\left[\left(\tilde{\mathbf{X}} - \mu_X\right)\left(\tilde{\mathbf{X}} - \mu_X\right)^T\right] \\
&= \mathbb{E}\left[\left(\Sigma_{XY}{\Sigma_Y}^{-1}(\mathbf{Y} - \mu_Y) + \mathbf{V}\right)\left(\Sigma_{XY}{\Sigma_Y}^{-1}(\mathbf{Y} - \mu_Y) + \mathbf{V}\right)^T\right] \\
&= \mathbb{E}\left[\left(\Sigma_{XY}{\Sigma_Y}^{-1}(\mathbf{Y} - \mu_Y)\right)\left(\Sigma_{XY}{\Sigma_Y}^{-1}(\mathbf{Y} - \mu_Y)\right)^T\right] + \mathbb{E}\left[\mathbf{V}\left(\Sigma_{XY}{\Sigma_Y}^{-1}(\mathbf{Y} - \mu_Y)\right)^T\right] \\
&\quad + \mathbb{E}\left[\mathbf{V}^T\Sigma_{XY}{\Sigma_Y}^{-1}(\mathbf{Y} - \mu_Y)\right] + \mathbb{E}\left[\mathbf{V}\mathbf{V}^T\right] \\
&= \mathbb{E}\left[\Sigma_{XY}{\Sigma_Y}^{-1}(\mathbf{Y} - \mu_Y)(\mathbf{Y} - \mu_Y)^T\Sigma_Y{}^{-1}\Sigma_{YX}\right] + \mathbb{E}[\mathbf{V}]\mathbb{E}\left[\Sigma_{XY}{\Sigma_Y}^{-1}(\mathbf{Y} - \mu_Y)^T\right] \\
&\quad + \mathbb{E}\left[\mathbf{V}^T\right]\mathbb{E}\left[\Sigma_{XY}{\Sigma_Y}^{-1}(\mathbf{Y} - \mu_Y)\right] + \mathbb{E}\left[\mathbf{V}\mathbf{V}^T\right] \\
&= \Sigma_{XY}{\Sigma_Y}^{-1}\mathbb{E}\left[(\mathbf{Y} - \mu_Y)(\mathbf{Y} - \mu_Y)^T\right]\Sigma_Y{}^{-1}\Sigma_{YX} + \Sigma_V \\
&= \Sigma_{XY}{\Sigma_Y}^{-1}\Sigma_{YX} + \Sigma_X - \Sigma_{XY}{\Sigma_Y}^{-1}\Sigma_{YX} \\
&= \Sigma_X
\end{aligned}$$

Thus, $\tilde{\mathbf{X}}$ and $\mathbf{Y}$ are jointly gaussian with mean $\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$ and covariance $\begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}$. This indicates that $(\tilde{\mathbf{X}}, \mathbf{Y}) \overset{d}{=} (\mathbf{X}, \mathbf{Y})$.

Theorem 24.2

For $\mathbf{X}, \mathbf{Y}$ jointly gaussian,

$$\begin{aligned}\mathbb{E}[\mathbf{X} \mid \mathbf{Y}] &= \mathbb{E}[\boldsymbol{\mu}_{\mathbf{X}} + \boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{Y}}(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}}) + \mathbf{V} \mid \mathbf{Y}] \\ &= \boldsymbol{\mu}_{\mathbf{X}} + \boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{Y}}{}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}}) \\ &= \mathbb{L}[\mathbf{X} \mid \mathbf{Y}]\end{aligned}$$

The linear least squares estimator coincides with the optimal minimum mean square error estimator for joint gaussians.

In practice, things are often approximately gaussian (by CLT). So, we can expect linear estimators in these instances to perform near-optimally.

25 Lecture 25

25.1 Jointly Gaussians Random Variables Cont.

There are many equivalent ways of characterizing jointly gaussian vectors, some of which include but are not limited to

- Specifying the density
- Affine transformations of IID $\mathcal{N}(0,1)$ random variables
- All one dimensional projections are single dimensional gaussian random variables
- Maximum entropy distribution subject to 2nd moment constraints
- Limit in the CLT
- If $\mathbf{X}, \mathbf{Y}$ are jointly Gaussian, then $\mathbf{X} = \boldsymbol{\mu}_{\mathbf{X}} + \boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{Y}}{}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}}) + \mathbf{V}$ for $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{Y}}{}^{-1}\boldsymbol{\Sigma}_{\mathbf{YX}})$. In other words, in a gaussian vector, each coordinate is a noisy linear combination of the others.
- Linear estimation is the most optimal form of estimation for gaussians
- Uncorrelated is equivalent to independent

Note 25.1

Gaussian marginals don't imply jointly gaussian. For example, take $Y \sim \mathcal{N}(0,1)$ and let

$$B = \begin{cases} +1 & \text{w.p. } \frac{1}{2} \\ -1 & \text{w.p. } \frac{1}{2} \end{cases}$$

be independent of Y. Then, $X = BY \sim \mathcal{N}(0,1)$ but (X,Y) are not jointly Gaussian. The probability density will be concentrated on the lines $x = y$ and $x = -y$ instead of a regular jointly gaussian curve/surface.

25.2 Kalman Filtering

Basic setting: let $\mathbf{X}_0, \mathbf{V}_0, \mathbf{V}_1, \ldots, \mathbf{W}_0, \mathbf{W}_1, \ldots$ be uncorrelated random vectors, say zero-mean (WLOG). The state space model is an evolution of the form

$$\begin{aligned} \mathbf{X}_{n+1} &= \mathbf{A}\mathbf{X}_n + \mathbf{V}_n \quad n \geq 0 \\ \mathbf{Y}_n &= \mathbf{C}\mathbf{X}_n + \mathbf{W}_n \quad n \geq 1 \end{aligned}$$

where $\mathbf{A}$ and $\mathbf{C}$ are assumed to be known matrices. This is a really flexible model for a variety of processes.

Note 25.2

If $\mathbf{X}_0, \mathbf{V}_0, \mathbf{V}_1, \ldots, \mathbf{W}_0, \mathbf{W}_1, \ldots$ are gaussians, then everything is jointly gaussian.

Example 25.1

Motion of a vehicle: let $p(n)$ be the position at time n. Then

$$\mathbf{X}_n = \begin{bmatrix} p(n) \\ p(n-1) \\ p(n-2) \end{bmatrix}$$

$$\mathbf{X}_{n+1} = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{X}_n + \begin{bmatrix} Z_n \\ 0 \\ 0 \end{bmatrix}, \quad Z_n \sim \mathcal{N}(0, \sigma^2)$$

$$\mathbf{Y}_n = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \mathbf{X}_n + W_n, \qquad W_n \sim \mathcal{N}(0, \xi^2)$$

Definition 25.1: Kalman Filter
The Kalman Filter is an efficient algorithm for estimating the X process sequentially from the observations Y.

There are many variations of it possible:

- Prediction - estimating X_{n+k} from $Y_1, \ldots, Y_n$
- Filtering - estimating X_n from $Y_1, \ldots, Y_n$
- Smoothing - estimating X_{n-k} from $Y_1, \ldots, Y_n$

Let $(\mathbf{X}_n)_{n \geq 0}$ evolve according to the state space model given above. For notational convenience, let $\hat{\mathbf{X}}_{n|m}$ denote $\mathbb{L}[\mathbf{X}_n \mid \mathbf{Y}^m]$, $\mathbf{\Sigma_{n|m}}$ denote $\mathrm{Cov}(\mathbf{X}_n - \hat{\mathbf{X}}_{n|m})$ (the covariance of the estimation error of $\mathbf{X}_n$ given $\mathbf{Y}^n$), $\mathbf{\Sigma_V}$ denotes $\mathrm{Cov}(\mathbf{V}_i)$ and $\mathbf{\Sigma_W}$ denote $\mathrm{Cov}(\mathbf{W}_i)$ for $i \geq 0$ (assume that all $\mathbf{V}_i$ and $\mathbf{W}_i$ have the same covariance for simplicity).

Theorem 25.1: Kalman Filter
Initialize: $\hat{\mathbf{X}}_{0|0} = \mathbf{0}$, $\mathbf{\Sigma_{0|0}} = \mathrm{Cov}(\mathbf{X}_0)$. For $n \geq 1$, do

$$\begin{aligned}
\hat{\mathbf{X}}_{n|n-1} &= \mathbf{A}\hat{\mathbf{X}}_{n-1|n-1} \\
\hat{\mathbf{Y}}_n &= \mathbf{Y}_n - \mathbb{L}\left[\mathbf{Y}_n \mid \mathbf{Y}^{n-1}\right] \\
&= \mathbf{Y}_n - \mathbb{L}\left[\mathbf{C}\mathbf{X}_n + \mathbf{W}_n \mid \mathbf{Y}^{n-1}\right] \\
&= \mathbf{Y}_n - \mathbf{C}\hat{\mathbf{X}}_{n|n-1} \\
\hat{\mathbf{X}}_{n|n} &= \hat{\mathbf{X}}_{n|n-1} + \mathbf{k}_n \hat{\mathbf{Y}}_n \\
&= \mathbf{A}\hat{\mathbf{X}}_{n-1|n-1} + \mathbf{K}_n\left(\mathbf{Y}_n - \mathbf{C}\mathbf{A}\hat{\mathbf{X}}_{n-1|n-1}\right) \\
\mathbf{K}_n &= \mathbf{\Sigma_{n|n-1}}\mathbf{C}^T\left(\mathbf{C}\mathbf{\Sigma_{n|n-1}}\mathbf{C}^T + \mathbf{\Sigma_W}\right)^{-1} \\
\mathbf{\Sigma_{n|n-1}} &= \mathbf{A}\mathbf{\Sigma_{n-1|n-1}}\mathbf{A}^T + \mathbf{\Sigma_V} \\
\mathbf{\Sigma_{n|n}} &= (\mathbf{I} - \mathbf{K}_n\mathbf{C})\mathbf{\Sigma_{n|n-1}}
\end{aligned}$$

25.3 Proof of correctness of the scalar KF

Consider the scalar example where

$$\begin{aligned}
X_n &= aX_{n-1} + V_n \\
Y_n &= X_n + W_n
\end{aligned}$$

for $n \geq 1$ and X_0 is zero-mean. The Kalman Filter of this model is given by

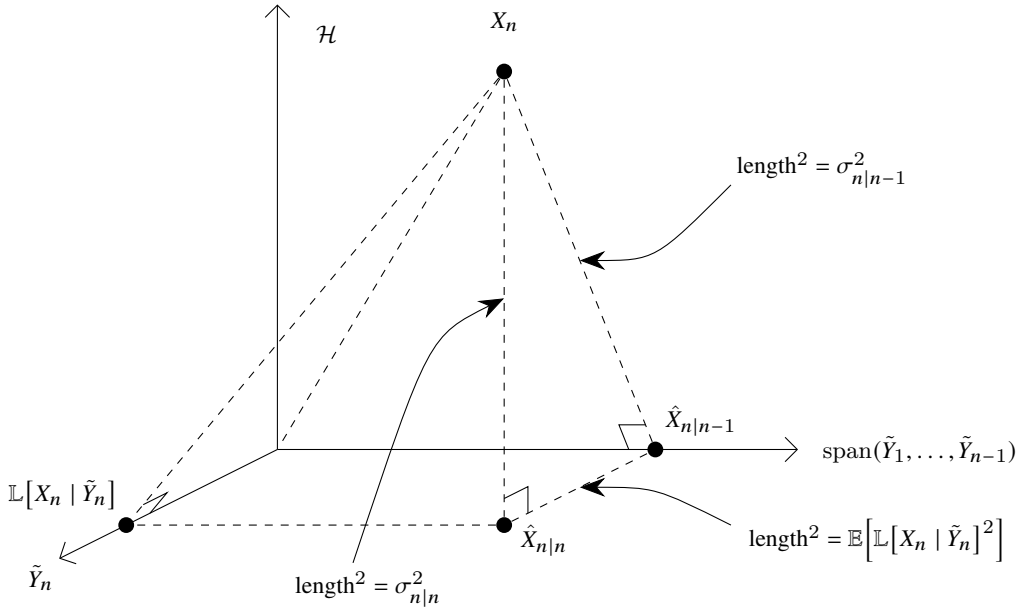
1. Initialize: $\hat{X}_{0|0}$, $\sigma^2_{0|0} = \mathrm{Var}(X_0)$
2. Updates:

$$\begin{aligned}
\hat{X}_{n|n} &= a\hat{X}_{n-1|n-1} + K_n\left(Y_n - a\hat{X}_{n-1|n-1}\right) \\
K_n &= \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_W} \\
\sigma^2_{n|n-1} &= a^2\sigma^2_{n-1|n-1} + \sigma^2_V \\
\sigma^2_{n|n} &= (1 - K_n)\sigma^2_{n|n-1}
\end{aligned}$$

The initialization is trivially correct. So, it suffices to verify the update equations. Let $\tilde{Y}_1, \tilde{Y}_2, \ldots$ be the linear innovation sequence corresponding to $Y_1, Y_2, \ldots$ (follows from Gram-Schmidt since $\tilde{Y}_n = Y_n - \mathbb{L}\left[Y_n \mid \mathbf{Y}^{n-1}\right]$ for $n \geq 1$). Then,

$$\begin{aligned}
\hat{X}_{n|n} &= \mathbb{L}\left[X_n \mid \mathbf{Y}^n\right] \\
&= \mathbb{L}\left[X_n \mid \tilde{\mathbf{Y}}^n\right] \\
&= \mathbb{L}\left[X_n \mid \tilde{\mathbf{Y}}^{n-1}\right] + \mathbb{L}\left[X_n \mid \tilde{Y}_n\right] \\
&= \hat{X}_{n|n-1} + \underbrace{\frac{\operatorname{Cov}(X_n, \tilde{Y}_n)}{\operatorname{Var}(\tilde{Y}_n)}}_{K_n} \tilde{Y}_n
\end{aligned}$$

Consider the Hilbert space of random variables and two subspaces given by $\operatorname{span}(\tilde{Y}_1, \ldots, \tilde{Y}_{n-1})$ and $\operatorname{span}(\tilde{Y}_n)$. Geometrically,



By looking at the figure above, and the properties of LLSE, we have

$$\begin{aligned}
\sigma^2_{n|n} &= \sigma^2_{n|n-1} - \mathbb{E}\left[\left(\mathbb{L}\left[X_n \mid \tilde{Y}_n\right]\right)^2\right] \\
&= \sigma^2_{n|n-1} - \mathbb{E}\left[\operatorname{Cov}(X_n, \tilde{Y}_n) \operatorname{Var}(\tilde{Y}_n)^{-1} \tilde{Y}_n \tilde{Y}_n^T \operatorname{Var}(\tilde{Y}_n)^{-1} \operatorname{Cov}(\tilde{Y}_n, X_n)\right] \\
&= \sigma^2_{n|n-1} - \operatorname{Cov}(X_n, \tilde{Y}_n) \operatorname{Var}(\tilde{Y}_n)^{-1} \operatorname{Var}(\tilde{Y}_n) \operatorname{Var}(\tilde{Y}_n)^{-1} \operatorname{Cov}(X_n, \tilde{Y}_n) \\
\sigma^2_{n|n} &= \sigma^2_{n|n-1} - \frac{\operatorname{Cov}(X_n, \tilde{Y}_n)^2}{\operatorname{Var}(\tilde{Y}_n)} \\
&= \sigma^2_{n|n-1} - K_n \operatorname{Cov}(X_n, \tilde{Y}_n)
\end{aligned}$$

Now, let's estimate

$$\begin{aligned}
\operatorname{Cov}(X_n, \tilde{Y}_n) &= \operatorname{Cov}\left(X_n, Y_n - \mathbb{L}\left[Y_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right) \\
&= \operatorname{Cov}\left(X_n, Y_n - \mathbb{L}\left[X_n + W_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right) \\
&= \operatorname{Cov}\left(X_n, Y_n - \mathbb{L}\left[X_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right), \quad W_n \text{ is uncorrelated to } \tilde{\mathbf{Y}}^{n-1} \\
&= \operatorname{Cov}\left(X_n, X_n + W_n - \mathbb{L}\left[X_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right) \\
&= \operatorname{Cov}\left(X_n, X_n - \mathbb{L}\left[X_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right), \quad W_n \text{ is uncorrelated to } X_n
\end{aligned}$$

$$= \operatorname{Cov}\left(X_n, X_n - \mathbb{L}\left[X_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right) - \underbrace{\operatorname{Cov}\left(\mathbb{L}\left[X_n \mid \tilde{\mathbf{Y}}^{n-1}\right], X_n - \mathbb{L}\left[X_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right)}_{=0 \text{ by the orthogonality principle}}$$

The above holds since LLSE error is orthogonal to any linear function of the observations and $\mathbb{L}\left[\cdot \mid \tilde{\mathbf{Y}}^{n-1}\right]$ is one. Then,

$$\begin{aligned}
\operatorname{Cov}(X_n, \tilde{Y}_n) &= \operatorname{Cov}\left(X_n - \mathbb{L}\left[X_n \mid \tilde{\mathbf{Y}}^{n-1}\right], X_n - \mathbb{L}\left[X_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right) \\
&= \operatorname{Var}\left(X_n - \mathbb{L}\left[X_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right) \\
&= \sigma^2_{n|n-1}
\end{aligned}$$

Thus,

$$\begin{aligned}
\sigma^2_{n|n} &= \sigma^2_{n|n-1} - K_n \operatorname{Cov}(X_n, \tilde{Y}_n) \\
&= \sigma^2_{n|n-1}(1 - K_n)
\end{aligned}$$

Note that

$$\begin{aligned}
\hat{X}_{n|n} &= \hat{X}_{n|n-1} + K_n\left(Y_n - \mathbb{L}\left[Y_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right) \\
&= \mathbb{L}\left[aX_{n-1} + V_{n-1} \mid \tilde{\mathbf{Y}}^{n-1}\right] + K_n\left(Y_n - \mathbb{L}\left[X_n + W_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right) \\
&= \mathbb{L}\left[aX_{n-1} \mid \tilde{\mathbf{Y}}^{n-1}\right] + K_n\left(Y_n - \mathbb{L}\left[aX_{n-1} \mid \tilde{\mathbf{Y}}^{n-1}\right]\right) \\
&= a\mathbb{L}\left[X_{n-1} \mid \tilde{\mathbf{Y}}^{n-1}\right] + K_n\left(Y_n - a\mathbb{L}\left[X_{n-1} \mid \tilde{\mathbf{Y}}^{n-1}\right]\right) \\
&= a\hat{X}_{n-1|n-1} + K_n\left(Y_n - a\hat{X}_{n-1|n-1}\right)
\end{aligned}$$

Moreover,

$$\begin{aligned}
\sigma^2_{n|n-1} &= \mathbb{E}\left[\left(X_n - \mathbb{L}\left[X_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(aX_{n-1} + V_n - \mathbb{L}\left[aX_{n-1} + V_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right)^2\right] \\
&= \mathbb{E}\left[a^2\left(X_{n-1} - \mathbb{L}\left[X_{n-1} + \mid \tilde{\mathbf{Y}}^{n-1}\right]\right)^2\right] + \mathbb{E}\left[V_n^2\right] \\
&= a^2\sigma^2_{n-1|n-1} + \sigma^2_V
\end{aligned}$$

And,

$$\begin{aligned}
\operatorname{Var}(\tilde{Y}_n) &= \mathbb{E}\left[\left(Y_n - \mathbb{L}\left[Y_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right)\right] \\
&= \mathbb{E}\left[\left(X_n + W_n - \mathbb{L}\left[X_n + W_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(X_n - \mathbb{L}\left[X_n \mid \tilde{\mathbf{Y}}^{n-1}\right]\right)^2\right] + \mathbb{E}\left[W_n^2\right] \\
&= \sigma^2_{n|n-1} + \sigma^2_W
\end{aligned}$$

Thus,

$$K_n = \frac{\operatorname{Cov}(X_n, \tilde{Y}_n)}{\operatorname{Var}(\tilde{Y}_n)} = \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_W}$$

Thus, we have verified all of the Kalman filter equations for the scalar case. The vector case is just a generalization of this.