Contents

1 Elementary Probability

1.1 Lecture 1

This lecture I did not have a laptop yet, so I was unable to transcribe anything. Here is what I remember was discussed:

Definition 1.1 (Conditional Probability) For two events A, B, the probability of A given B is:

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

Definition 1.2 (MAP) We say the Most likely A Posteriori (MAP) estimate of a random variable X given $Y = y$, is:

$$\operatorname*{argmax}_x \mathbb{P}[X = x \mid Y = y]$$

Definition 1.3 (MLE) We say the Maximum Likelihood Estimate (MLE) of a random variable X given Y, is:

$$\operatorname*{argmax}_x \mathbb{P}[Y = y \mid X = x]$$

1.2 Lecture 2

1.2.1 Probability Fundamentals and Random Variables

We begin probability by defining a set Ω called the sample space. Elements of the sample space are termed outcomes. Subsets of Ω are termed as events.

For some event A, we can define the probability of A as follows:

$$\mathbb{P}[A] = \sum_{\omega \in A} \mathbb{P}[\omega]$$

Probability maps events to $[0, 1]$ in a consistent manner, satisfying the following axioms:

- $\mathbb{P}[\Omega] = 1$
- $\mathbb{P}[\emptyset] = 0$
- For two disjoint events A_1, A_2, we have $\mathbb{P}[A_1 \cup A_2] = \mathbb{P}[A_1] + \mathbb{P}[A_2]$

This is what we term a probability space. Often it is more helpful to work with events than with individual sample points (especially in the case of an uncountably infinite amount of sample points).

Definition 1.4 (Random Variable) A random variable $X : \omega \to B$ maps each outcome to elements of some other set (often $\mathbb{R}$). $X = x$ for some x is an event, with a well-defined probability.

Definition 1.5 (Independence) Two random variables X and Y are independent if

$$\mathbb{P}[X = x | Y = x] = \mathbb{P}[X = x]$$

i.e.

$$\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x]\,\mathbb{P}[Y = y]$$

Example 1.1 Suppose you flip a coin 10 times. We will show that X, the amount of heads in the first 4 flips, and Y, the amount of heads in the last 6 flips, are independent.
Let $a(x)$ be the amount of ways to get x heads in 4 flips and $b(y)$ be the amount of ways to get y heads in 6 flips. Then,

$$\mathbb{P}[X = x] = \frac{a(x) \cdot 2^6}{2^{10}} = \frac{a(x)}{2^4}$$
$$\mathbb{P}[Y = y] = \frac{b(y) \cdot 2^4}{2^{10}} = \frac{b(y)}{2^6}$$
$$\mathbb{P}[X = x, Y = y] = \frac{a(x) \cdot b(y)}{2^{10}} = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y]$$

Thus, the random variables are independent.

Definition 1.6 (Expectation) The expectation of a (discrete) random variable is:

$$\mathbb{E}[X] = \sum_x x \mathbb{P}[X = x]$$

This is often called the mean or the average value.

Theorem 1.1 Properties of expectation:

- $\mathbb{E}[a] = a$ for $a \in \mathbb{R}$
- If the space is uniform, then $\mathbb{E}[X] = \frac{1}{N}\sum_x x$
- $\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$ for $\alpha, \beta \in \mathbb{R}$
- $X \leq Y \Rightarrow \mathbb{E}[X] \leq \mathbb{E}[Y]$
- If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \not\Rightarrow X, Y$ independent

There are two ways of thus computing expectation. You can either sum over sample points, or take a lot of measurements of your random variable, then divide by the amount of measurements. The reason this works is because of property 2 above.

Definition 1.7 (Variance and Standard Deviation) The variance of a random variable X is defined as:

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X]]^2$$

The standard deviation of this random variable is:

$$\sigma_X = \sqrt{\text{Var}(X)}$$

The variance measures the spread away from the mean that a random variable may exhibit.

Theorem 1.2 Properties of variance:

- $\text{Var}(X) \geq 0$, with equality only if X is constant
- $\text{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$
- $\text{Var}(aX) = a^2\text{Var}(X)$ for constant a
- If X and Y are independent, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$
- In general, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$

Often we term $\mathbb{E}\left[X^k\right]$ as the kth moment of X, so the variance contains information about the second moment of X.

1.3 Lecture 3

1.3.1 Concentration Inequalities

Definition 1.8 (Indicator Random Variable) $\mathbb{1}_A$ is the indicator for event A, i.e. a random variable with the following values:

$$\mathbb{1}_A = \begin{cases} 1 & \text{if sample point in event } A \\ 0 & \text{otherwise} \end{cases}$$

Theorem 1.3 (Markov's Inequality) Consider random variable $X \geq 0$ and constant $a > 0$. Then,

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

Proof Let $Y = \mathbb{1}_{X \geq a}$. Then we know:

$$\begin{aligned} \mathbb{E}[Y] &= 0 \cdot \mathbb{P}[X < a] + 1 \cdot \mathbb{P}[X \geq a] = \mathbb{P}[X \geq a] \\ Y &\leq \frac{X}{a} \\ \mathbb{E}[Y] &\leq \mathbb{E}\left[\frac{X}{a}\right] = \frac{\mathbb{E}[X]}{a} \\ \mathbb{P}[X \geq a] &\leq \frac{\mathbb{E}[X]}{a} \end{aligned}$$

Markov's inequality tends to be a coarse bound, and X, a have to be non-negative.

Theorem 1.4 (Chebyshev's Inequality) For random variable X and $\epsilon > 0$:

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

Define $Z = |X - \mathbb{E}[X]|^2$, $a = \epsilon^2$, $\epsilon > 0$.

Proof Apply Markov's inequality:

$$\mathbb{P}\left[Z \geq \epsilon^2\right] \leq \frac{\mathbb{E}[Z]}{\epsilon^2}$$

Note that

$$\mathbb{E}[Z] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathrm{Var}(X)$$

This means that

$$\mathbb{P}\left[\sqrt{Z} \geq \epsilon\right] \leq \frac{\mathrm{Var}(X)}{\epsilon^2}$$

which is exactly the statement of Chebyshev's.

Chebyshev's is generally a tighter bound than Markov's.

Theorem 1.5 (Weak Law of Large Numbers) Assume $X_1, X_2, X_3, \ldots$ are independent random variables with the same expectation μ and the same variance σ^2, and define $Y_n = \frac{(X_1 + X_2 + \cdots + X_n)}{n}$ Then, we have that for any

constant $\epsilon > 0$.

$$\lim_{n\to\infty} \mathbb{P}\left[|Y_n - \mu| \geq \epsilon\right] \to 0$$

This can be shown by Chebyshev's inequality, namely note that the expression in the limit is bounded by $\frac{\text{Var}(Y_n)}{\epsilon^2} = \frac{n\sigma^2}{n^2\epsilon^2} \to 0$.
In words, this means the probability of the sample mean being within ϵ of the true mean approaches 1.

1.3.2 Covariance and Estimation

Definition 1.9 (Covariance)

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

If X and Y are independent, then $\text{Cov}(X,Y) = 0$. If the latter is true, then X and Y are uncorrelated. We also define the coefficient of correlation:

$$\rho_{xy} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

This is handy because $|\rho_{XY}| \leq 1$.

Suppose we want to estimate a random variable Y by $\hat{Y}$ given a correlated random variable Y.

We want to minimize $\mathbb{E}\left[\left(Y - \hat{Y}\right)^2\right]$ but have a linear relationship. This yields LLSE:

Theorem 1.6 (Linear Least Squares Estimate) The LLSE

$$\hat{Y} = \mathbb{E}[Y] + \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(X - \mathbb{E}[X])$$

is the best linear estimate of Y given X.

2 Basic Probability

2.1 Lecture 3, Continued

2.1.1 Infinite Collections of Events and Borel-Cantelli

There are some important consequences of the axioms of probability.

Theorem 2.1 (Infinite Sub-Events) Consider some set A where $A = \bigcup_{n=1}^{\infty} A_n$ where:

$$A_1 \subseteq A_2 \subseteq \dots$$

Then, $\mathbb{P}[A_n] \to \mathbb{P}[A]$.
Furthermore, consider some set B where $B = \bigcap_{n=1}^{\infty} B_n$ where:

$$B_1 \supseteq B_2 \supseteq \dots$$

Then, $\mathbb{P}[B_n] \to \mathbb{P}[B]$.

Theorem 2.2 (Borel-Cantelli Theorem) Let $\{A_n\}_{n=1}^{\infty}$ be a collection of events such that $\sum_{n=1}^{\infty} \mathbb{P}[A_n] < \infty$, then $\mathbb{P}[A_n \text{ infinitely often}] = 0$.
$\{A_n \text{ infinitely often}\}$ is the following event:

$$\{\omega \mid \exists N(\omega) \text{ such that } \forall n > N(\omega), \omega \notin A_n\}$$

In English, the set describes all outcomes where you can assign a number N to that outcome such that after A_N, you have no set membership in a greater A_n.
The theorem claims that you CAN assign such a number (max event) to any outcome with nonzero probability.

Proof Suppose we have such a sequence of A_n such that $\sum_{n=1}^{\infty} \mathbb{P}[A_n] < \infty$. Define

$$B_n = \bigcup_{m \geq n} A_m$$

Note that $B_1 \supseteq B_2 \supseteq B_3 \ldots$
Calling,

$$B = \bigcap_{n} B_n$$

Note that $\omega \in \{A_n \text{ i.o.}\}$ if and only if $\omega \in B_n$ for all n. But this means that $\omega \in B$. This means that $\{A_n \text{ i.o.}\} = B$. So we must calculate $\mathbb{P}[B]$. However, note that $\mathbb{P}[B_n] \to \mathbb{P}[B]$ as $n \to \infty$. Thus, we must compute:

$$\begin{aligned}
\mathbb{P}[B] &= \lim_{n \to \infty} \mathbb{P}[B_n] \\
\mathbb{P}[B_n] &\leq \sum_{m=n}^{\infty} \mathbb{P}[A_m] \to 0 \\
\mathbb{P}[B] &= 0
\end{aligned}$$

The second step is justified by the union bound and following result from analysis. For non-negative sequence a_n, if $\sum_{i=1}^{\infty} a_n < \infty$, then $\lim_{n \to \infty} \sum_{m=n}^{\infty} a_m \to 0$.
Our result shows that $\mathbb{P}[A_n \text{ i.o.}] = 0$ ∎

Consider the following example for coin flips:

Example 2.1 (Infinite Coin Flips) Consider the experiment of flipping a coin infinitely many times. Let

$$A_n = \{n\text{th flip is heads}\}$$

Then, in this experiment, the event $\{A_n \text{ infinitely often }\}$ (which we denote as $\{A_n \text{ io}\}$)

$$\{A_n \text{ io}\} = \{\omega \mid \text{heads never stop after some } N(\omega)\}$$

Here are some sequences that are in that event:

$$\begin{aligned}
\omega &= 0, 0, 1, 1, 1, 1, \ldots \\
\omega &= 0, 1, 0, 1, 0, 1, \ldots \\
\omega &= 0, 0, \underbrace{\ldots}_{\text{1 million 0's}}, 1, 0, 0, \underbrace{\ldots}_{\text{1 million 0's}}, 1, \ldots
\end{aligned}$$

Now consider the assigning the following probabilities to each heads (instead of the normal, uniform probability space):

$$\mathbb{P}[A_n] = \frac{1}{n^2}$$

$\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges, so by Borel-Cantelli, $\mathbb{P}[A_n \text{ io}] = 0$, i.e. the heads ALWAYS stop.
Now there is one more question. Does $\mathbb{P}[A_n \text{ i.o.}] = 0 \implies \{A_n \text{i.o.}\} = 0$? The answer is no. In this case, $\mathbb{P}[A_n \text{i.o.}] = 0$, but consider the outcome ω_n where the nth flip onwards is a heads; these are all in the infinitely often set, so it actually has infinite cardinality!

2.2 Lecture 4

2.2.1 The Laws of Large Numbers, Revisited

Here is a recap of the two different laws of large numbers.

First, we define two different types of convergence:

Definition 2.1 (Almost Sure Convergence) A random variable X_n almost surely converges to random variable X if

$$\mathbb{P}\left[X_n \to X\right] = 1$$

as $n \to \infty$.

Definition 2.2 (Convergence in Probability) A random variable X_n converges in probability to random variable X if

$$\mathbb{P}\left[|X_n - X| > \epsilon\right] \to 0$$

for any real number $\epsilon > 0$ as $n \to \infty$.

Theorem 2.3 (Strong Law of Large Numbers) Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (iid) random variables. Define:

$$Y_n = \frac{X_1 + \cdots + X_n}{n}$$
$$Y = \mathbb{E}\left[X_1\right]$$

Y_n converges to Y almost surely.

Note the contrast with the weak law of natural numbers. The weak law had only convergence in probability. A key thing to note is that the strong law **implies** the weak law.

2.2.2 Independence

Let us now refine the notion of Independence.

Definition 2.3 (Pairwise Independence) Consider events A_j with $j \in J$.
The events are pairwise independent if for any $j, k \in J$,

$$\mathbb{P}\left[A_j \cap A_k\right] = \mathbb{P}\left[A_j\right]\mathbb{P}\left[A_k\right]$$

Definition 2.4 (Mutual Independence) Consider events A_j with $j \in J$.
The events are mutually independent if

$$\mathbb{P}\left[\bigcap_{j \in K} A_j\right] = \prod_{j \in K} \mathbb{P}\left[A_j\right], \forall K \subseteq J$$

Note that pairwise independence does not imply mutual independence. Here is an example of that edge case:

Example 2.2 Take probability space $\Omega = \{1, 2, 3, 4\}$, all equally likely. Consider the events: $A = \{1, 2\}, B = \{1, 3\}, C = \{1,4\}$.
Note that $\mathbb{P}[A \cap B] = \frac{1}{4} = \mathbb{P}[A]\mathbb{P}[B]$,
but $\mathbb{P}[A \cap B \cap C] = \frac{1}{4} \neq \frac{1}{8} = \mathbb{P}[A]\mathbb{P}[B]\mathbb{P}[C]$.

Now with independence, we can find that the converse of Borel-Cantelli is often true:

Theorem 2.4 (Converse of Borel-Cantelli Theorem) Let A_n be a collection of mutually indpendent events such that $\sum_{n=1}^{\infty} \mathbb{P}[A_n] = \infty$. Then, $\mathbb{P}[A_n \text{ infinitely often}] = 1$.

Let us use another example to understand this:

Example 2.3 Let A_n be the same event as the other example (the nth flip is heads) and assign:

$$\mathbb{P}[A_n] = \frac{1}{n}$$

where all the A_n are mutually independent.
Since $\sum_{n=1}^{\infty} \frac{1}{n} = \infty$ and thus by the converse of Borel-Cantelli: $\mathbb{P}[A_n \text{ io}] = 1$.

Example 2.4 (Glued Coins) Suppose you have n coins that are all glued together, i.e. the only two outcomes are $HHH\ldots$ or $TTT\ldots$. Then let A_n = the nth coin is heads. Note that

$$\mathbb{P}[A_n] = \frac{1}{2}$$

which means $\sum \mathbb{P}[A_n] = \infty$. Thus the probability of getting a sequence which has heads infinitely often is 1.

Theorem 2.5 (Kolmogorov's 0-1 theorem) If you have a set of events $\{A_n\}_{n=1}^{\infty}$ that all independent, then

$$\mathbb{P}[A_n \text{ infinitely often}] = 0 \text{ or } 1$$

2.2.3 Conditional Probability

Now we refine conditional probability for many events.

Definition 2.5 (Conditional Probability) Let A and B be two events, and assume $\mathbb{P}[B] > 0$. Then the conditional probability of A given B is:

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

Theorem 2.6 (Chain Rule) For two events we had $\mathbb{P}[A \cap B] = \mathbb{P}[A \mid B]\mathbb{P}[B]$. For n events A_i, we have:

$$\mathbb{P}[A_1 \cap A_2 \cap \cdots \cap A_n] = \mathbb{P}[A_1]\mathbb{P}[A_2 \mid A_1]\mathbb{P}[A_3 \mid A_1 \cap A_2]\ldots\mathbb{P}[A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}]$$

if $\mathbb{P}[A_1 \cap A_2 \cap \ldots A_{n-1}] > 0$.

The generalized result above can be shown by induction, taking the case of two events as the base case and then inducting on n. Now we will bring in some of the most powerful tools.

Theorem 2.7 (Law of Total Probability) Let $A, B_1, \ldots B_n$ be events where B_i's are disjoint and $\bigcup_{i=1}^n B_i = \Omega$. Then,

$$\mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A \cap B_i]$$

Theorem 2.8 (Bayes' Rule) Let $A, B_1, \ldots B_n$ be events where B_i's are disjoint and $\bigcup_{i=1}^n B_i = \Omega$.

$$\mathbb{P}[B_i \mid A] = \frac{\mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]}{\sum_{j=1}^n \mathbb{P}\left[A \mid B_j\right]\mathbb{P}\left[B_j\right]}$$

Proof Note that we can use the initial definition to expand the left side:

$$\begin{aligned}\mathbb{P}[B_i \mid A] &= \frac{\mathbb{P}[A \cap B_i]}{\mathbb{P}[A]} \\ &= \frac{\mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]}{\sum_{j=1}^n \mathbb{P}\left[A \cap B_j\right]} \\ &= \frac{\mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]}{\sum_{j=1}^n \mathbb{P}\left[A \mid B_j\right]\mathbb{P}\left[B_j\right]}\end{aligned}$$

where the summation in the denominator comes from the law of total probability.

Often the B_j's are termed the prior probabilities, and A is considered the posterior probability.

For an event $B \subseteq R$, $\mathbb{P}[X \in B] = \mathbb{P}\left[(X^{-1}(B))\right]$ where

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\}$$

.

We can define the following for a random variable to the reals.

Definition 2.6 (Cumulative Distribution Function (CDF)) The Cumulative Distribution Function $F_X(x)$ of random variable X is defined by:

$$F_X(x) = \mathbb{P}[X \in (-\infty, x]] = \mathbb{P}[X \leq x]$$

Here are some properties of the CDF:

- F_X is non-decreasing.
- F_X is right-continuous.
- $F_X \to 0$ as $x \to -\infty$ and $F_X \to 1$ as $x \to \infty$.

Example 2.5 (CDF of an Indicator) Consider the following random variable:

$$I = \begin{cases} 0 \text{ with probability } 1-p \\ 1 \text{ with probability } p \end{cases}$$

Then the $F_I(i)$ is a step function: TODO Add figure

2.3 Lecture 5

Definition 2.7 (Discrete Random Variable) A discrete random variable X can be described fully by:

$$\{(x_n, p_n), n = 1, \ldots, N\}$$

where $p_n = \mathbb{P}[X = x_n]$. This is called the probability mass function (PMF) of X.

We can write the expectation as follows:

$$\mathbb{E}[X] = \sum_{n=1}^{N} x_n p_n$$

With $N = \infty$, the expectation may not be defined.

Definition 2.8 (Function of a Random Variable) Calling $h(X)$ means changing to another random variable with the following PMF:

$$(h(x_n), p_n), n = 1, \ldots, N$$

The expectation of this is as follows:

$$\mathbb{E}[h(X)] = \sum_{n=1}^{N} h(x_n) p_n$$

Definition 2.9 (Coefficient of Variation) The coefficient of variation c of X is defined:

$$c = \sigma_X / \mathbb{E}[X]$$

2.3.1 Common Discrete Distributions

Bernoulli random variables model situations like individual coin flips.

Definition 2.10 (Bernoulli Random Variables) If $X =_D B(p)$ with $p \in [0, 1]$, then the PMF of X is:

$$\{(0, 1 - p), (1, p)\}$$

Furthermore, $\mathbb{E}[X] = p$ and $\text{Var}(X) = p(1 - p)$.

Geometric random variables model the situation where you count the number of coin flips until you get "heads".

Definition 2.11 (Geometric Random Variable) If $X =_D G(p)$ with $p \in [0, 1]$, then the PMF of X is:

$$\mathbb{P}[X = n] = (1 - p)^{n-1} p$$

Furthermore, $\mathbb{E}[X] = \frac{1}{p}$ and $\text{Var}(X) = \frac{1-p}{p^2}$.

The CDF can also be derived as $\mathbb{P}[X \leq n] = 1 - (1 - p)^n$, since it's the complement of failing n times. The CCDF (Complementary CDF) is thus $\mathbb{P}[X < n] = (1 - p)^n$.

Note 2.1 (Memoryless Property) The geometric distribution is memoryless, i.e. if $X =_D G(p)$, then

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

Binomial random variables model the situation of doing n coin flips and counting the heads, or the sum of n i.i.d. Bernoulli random variables.

Definition 2.12 (Binomial Random Variable) If $X =_D B(N, p)$ with $p \in [0, 1]$ and $N \geq 1$, then the PMF of X is:

$$\mathbb{P}[X = n] = \binom{N}{n} p^n (1-p)^{N-n}$$

Furthermore, $\mathbb{E}[X] = Np$ and $\text{Var}(X) = Np(1-p)$

The mode of the binomial distribution (the maximum probability) is at $n = \lfloor p(N+1) \rfloor$.

Poisson random variables are the limit of the binomials as the rate of coin flips goes to infinity. This represents the number of successes in an interval during a continuous process.

Definition 2.13 (Poisson Random Variable) If $X =_D P(\lambda)$ with $\lambda > 0$, then the PMF of X is:

$$\mathbb{P}[X = n] = \frac{e^{-\lambda}\lambda^n}{n!}$$

Furthermore, $\mathbb{E}[X] = \lambda$ and $\text{Var}(X) = \lambda$.

In fact, we can make this limit more precise.

Theorem 2.9 (Binomial Converges to Poisson) We have, setting $Np = \lambda$, where λ is fixed,

$$B(N, \lambda/N) \to P(\lambda)$$

2.3.2 Multiple Discrete Random Variables

Consider a pair of random variables (X, Y).

Definition 2.14 (Joint PMF) The joint distribution is given by:

$$p_{i,j} = \mathbb{P}\left[X = x_i, Y = y_j\right]$$

To find the PMF of one of the variables from the joint distribution, we can

Note 2.2 (Marginal PMF from JPMF)

$$\mathbb{P}[X = x_i] = \sum_j \mathbb{P}\left[X = x_i, Y = y_j\right]$$

Furthermore,

Theorem 2.10 (Independence for Random Variables) X and Y are independent if and only if

$$\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x]\,\mathbb{P}[Y = y]$$

If you have a function of multiple random variables, you can apply it similarly to the one variable case.

$$\mathbb{E}[h(X, Y)] = \sum_i \sum_j h(x_i, y_j)\mathbb{P}\left[X = x_i, Y = y_j\right]$$

First we extend the idea of conditioning to random variables.

Definition 2.15 (Conditional PMF) We call the conditional distribution of Y given X as:

$$\mathbb{P}\left[Y = y_j \mid X = x_i\right] = \frac{\mathbb{P}\left[X = x_i, Y = y_j\right]}{\mathbb{P}\left[X = x_i\right]}$$

Definition 2.16 (Conditional Expectation) The expectation of Y given X (i.e. the best guess of Y given X) is denoted $\mathbb{E}\left[Y \mid X\right]$ and is a function of X

Furthermore, if we want to use a function, we can compute it as follows:

$$\mathbb{E}\left[h(Y) \mid X = x_i\right] = \sum_j h(y_j)\mathbb{P}\left[Y = y_j \mid X = x_i\right]$$

Theorem 2.11 (Properties of Conditional Expectation) For two random variables X, Y,

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] &= \mathbb{E}\left[Y\right] \\
\mathbb{E}\left[h(X)Y \mid X\right] &= h(X)\mathbb{E}\left[Y \mid X\right] \\
\mathbb{E}\left[Y \mid X\right] &= \mathbb{E}\left[Y\right] \text{ if } X \text{ and } Y \text{ are independent} \\
\mathbb{E}\left[h_1(Y) + h_2(Y) \mid X\right] &= \mathbb{E}\left[h_1(Y) \mid X\right] + \mathbb{E}\left[h_2(Y) \mid X\right]
\end{aligned}$$

2.4 Lecture 6

Unfortunately, I didn't transcribe this lecture, as I was very tired. Here is one of the more important results.

In general $X_n \to X \not\Longrightarrow \mathbb{E}[X_n] \to \mathbb{E}[X]$. However, Dominated Convergence Theorem (DCT) and Monotone Convergence Theorem (MCT) provide sufficient conditions in the following form.

Theorem 2.12 (Continuous Tail Sum Formula) Let $X \geq 0$ be a non-negative random variable with $\mathbb{E}[X] < \infty$. Then,

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > x] \, \mathrm{d}x$$

2.5 Lecture 7

2.5.1 Multiple Continuous Random Variables

Definition 2.17 (JCDF) For random variables X and Y, the joint CDF (JCDF) is given by:

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

For continuous random variables, this is:

$$F_{X,Y}(x,y) = \int_{\infty}^{x}\int_{\infty}^{y} f_{X,Y}(x',y')\,\mathrm{d}y'\,\mathrm{d}x'$$

Then, note that to find $F_{\max X,Y}(k) = \mathbb{P}[\max X, Y \le k] = \mathbb{P}[X \le k, Y \le k]$, we can invoke the JCDF (and make simplifications if X and Y are independent). Furthermore, to find the probability density, we can simply differentiate with respect to k.

Similarly, $\mathbb{P}[\min X, Y > k] = \mathbb{P}[X > k, Y > k]$, which would be the JCCDF (which is the product of the CCDFs if X and Y are independent).

The sum of independent random variables is given by the convolution, *.

Theorem 2.13 (Convolution) Let $Z = X + Y$ where X and Y are independent. Then,

$$f_Z(z) = f_X * f_Y = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,\mathrm{d}x$$

In addition, we have conditioning for continuous random variables as well.

Definition 2.18 (Conditional PDF) Consider two random variables, X and Y such that $f_X(x) \neq 0$ at the point we are considering. Then:

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

Definition 2.19 (Conditional Variance) Let X and Y be random variables. Then we define conditional variance as:

$$\mathrm{Var}(Y \mid X) = \mathbb{E}\left[Y^2 \mid X\right] - (\mathbb{E}[Y \mid X])^2$$

Similar to the law of iterated expectation, there is a similar formula for variance:

Theorem 2.14 (Law of Iterated Variance) Let X and Y be random variables. Then,

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

Proof

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

An interpretation of this is to think about dividing Ω into disjoint sets where S_i where $X = x_i$. The conditional expectation, The first term is saying that taking the variance of replacing each set with the average over the set, is the first-order approximation. Then, to correct, we average the variances across each and add those in.

3 Page Rank

3.1 Lecture 7, Continued

Page rank is an algorithm originally used by Google for ranking the pages from a keyword search.

It tries to look at the problem as a Markov chain; the weight of each page p is sum over all pages linking to p, times the probability of visiting p from that other page. Symbolically, this is:

$$\pi(i) = \sum_{j \in \mathcal{X}} \pi(j) P(j,i), \forall i \in \mathcal{X}$$

where π is a (row) vector of the weights, and P is a matrix whose (j,i) entry corresponds to the probability of transitioning from j to i. So, we can rewrite the above equations as

$$\pi = \pi P$$

We also add the normalization constraint

$$\sum_{i \in \mathcal{X}} \pi(i) = 1$$

So, π is a probability distribution.

3.1.1 Discrete-Time Markov Chains

Definition 3.1 (Discrete Time Markov Chain (DTMC)) The DTMC $\{X(n), n \geq 0\}$ over state space $\mathcal{X}$ with $P = [P(i,j)]$ as the transition matrix with $P(i,j) = \mathbb{P}[X(n+1) = j \mid X(n) = i]$.
Markov chains have the memoryless property:

$$\mathbb{P}[X(n+1) = j \mid X(n) = i, X(m), m < n] = P(i,j) \forall i, j$$

Generally, we have P constant with respect to time, i.e. it's time-homogenous. Note that we have not shown that defining probabilities like this is consistent with the axioms of probability; let us assume such a choice exists for now.

Then, let $\pi_n(i)$ be the probability that the Markov chain is in state i at time n. Then, for every time step, using vector notation, we get the recurrence:

$$\pi_{n+1} = \pi_n P \implies \pi_n = \pi_0 P^n$$

Definition 3.2 (Stationary Distribution) Let P be the transition matrix of a markov chain. We call π a stationary distribution of the Markov chain if:

$$\pi = \pi P \text{ and } \sum_{i=1}^{n} \pi(i) = 1$$

We call these equations the balance equations.

One thing to note is that we can rewrite the equations as $(P - I)\pi = 0$. However, since every row sums to 0, this is not a full rank matrix, so we need the extra constraint about the sum of all the elements of π.

Next, we have some ways to classify Markov chains.

Definition 3.3 (Irreducible) A Markov Chain is irreducible if it can reach any state from any other state (possibly in multiple steps).

Definition 3.4 (Aperiodic) Let $d(i) = \gcd\{n \geq 1 \mid P^n(i,i) > 0\}$. An irreducible DTMC is aperiodic if $d(i) = 1$ for all i (in fact in an irreducible DTMC, $d(i)$ is the same as i).

Note that if a Markov chain has a self loop, it is aperiodic because we can get from i to i in a single step, so $d(i) = 1$ always.

Example 3.1 Consider the following Markov chain, which is irreducible and aperiodic. (DIAGRAM NEEDED)

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0.6 & 0 & 0.4 \\ 0 & 0.9 & 0.1 \end{bmatrix}$$

Note that $P(i,j)$ is the probability of going from i to j. Note that all the rows sum to 1 (by the total probability rule). Now let us solve the balance equations:

$$\pi = \pi P$$

$$\begin{bmatrix} \pi_0 & \pi_1 & \pi_2 \end{bmatrix} = \begin{bmatrix} \pi_0 & \pi_1 & \pi_2 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0.6 & 0 & 0.4 \\ 0 & 0.9 & 0.1 \end{bmatrix}$$

This yields the equations:

$$\begin{cases} 0.6\pi_1 = \pi_0 \\ \pi_0 + 0.9\pi_2 = \pi_1 \\ 0.4\pi_1 + 0.1\pi_2 = \pi_2 \\ \pi_0 + \pi_1 + \pi_2 = 0 \end{cases}$$

Note that some of these equations are dependent, so we needed the extra constraint. This can then be solved to get:

$$\pi = \begin{bmatrix} 0.294 & 0.489 & 0.217 \end{bmatrix}$$

With these notions, we have the following results.

Theorem 3.1 (Big Theorem for Finite DTMC) Consider an irreducible DTMC over a finite state space. Then,

- There is a unique invariant distribution π
- The long-term fraction of time spent in state n is given by

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{1}_{\{X(n)=i\}} = \pi(i)$$

- If the DTMC is aperiodic, $\pi_n \to \pi$ as $n \to \infty$, independent of π_0

Another way to state the last result is that each row in P^n converges to π as $n \to \infty$. Why? Consider $\pi_0 = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}$. Then $\pi_n = \pi_0 P^n = p_1 \to \pi$, so the first row approaches π. The argument is similar for any row.

3.2 Lecture 8

3.2.1 Hitting Times

Note 3.1 Consider a Markov Chain X_i. Suppose we wanted to find the time it takes to hit any state in some set A, starting at some other state i. Formally, we are asking,

$$\beta(i) := \mathbb{E}\left[T_A \mid X_0 = i\right], i \in \mathcal{X}, A \subseteq \mathcal{X}$$

Since Markov chains are memoryless, we only need to consider the ways to step from edges out of us, i.e. one step. We call these equations first-step equations (FSEs). The FSEs for this problem are:

$$\begin{cases} 1 + \sum_j P(i,j)\beta(j) & i \notin A \\ 0 & i \in A \end{cases}$$

The reasoning for this is as follows. If we are already in set A, it takes us 0 steps to get into it. Otherwise, we take one step to another state j, which has probability $P(i,j)$ of happening, and then the value of $\beta(i)$ will just be $1 + \beta(j)$.

This way of solving for hitting times produces n^2 linear equations to solve.

Here are some other problems we can solve with FSEs/Hitting times.

Note 3.2 Suppose you want to find the probability of hitting a state in one set A before another B,

$$\alpha(i) := \mathbb{P}\left[T_A < T_B \mid X_0 = i\right], i \in \mathcal{X}, A,B \subseteq \mathcal{X}, A \cap B = \emptyset$$

Then, the FSEs are the following:

$$\begin{cases} \sum_j P(i,j)\alpha(j) & i \notin A \cup B \\ 1 & i \in A \\ 0 & i \in B \end{cases}$$

The reasoning is similar: once you hit a bad state, your probability drops to 0 and if you hit a good state, your probability is 1. Otherwise, it is the sum of the probabilities of of moving to an adjacent state and that state ending up as something good (the sum is due to total probability rule).

Note 3.3 Consider a discounted reward for visiting states earlier. Define $h(i)$ as the reward for being in state i. Then we define

$$Z = \sum_{n=0}^{T_A} \beta^n h(X(n)), A \subseteq \mathcal{X}, 0 < \beta \leq 1$$

as the discounted reward (from a gambling run, perhaps). You accrue wealth until you hit A, but take too long and your reward is reduced. We want to find:

$$\mathbb{E}\left[Z \mid X_0 = i\right], i \in X$$

The FSEs then become:

$$\begin{cases} h(i) + \beta \sum_j P(i,j)\delta(j) & i \in A \\ h(i) & i \in A \end{cases}$$

3.3 Lecture 9

3.3.1 Laws of Large Numbers

We now discuss the law(s) of large numbers in detail, and how they relate to Markov chains.

Theorem 3.2 (Laws of Large Numbers) Let $\{X(i), i \geq 1\}$ be a sequence of independent and identically distributed (IID) random variables, with mean μ and let $S(n) = \sum_{i=1}^{n} X(i)$. Assume $\mathbb{E}[|X(i)|] < \infty$.
The Strong Law of Large Numbers (SLLN) states that:

$$\mathbb{P}\left[\lim_{n\to\infty} S(n)/n = \mu\right] = 1$$

i.e. $S(n)/n$, the sample mean, converges to the true mean μ **almost surely**.
The Weak Law of Large Numbers (WLLN) states that fixing $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{S(n)}{n} - \mu\right| > \varepsilon\right] = 0$$

i.e. $S(n)/n$, the same mean, converges to the true mean μ in **probability**.

Note that SLLN implies WLLN, so SLLN is stronger. Furthermore,

$$\mathbb{E}[S_n/n] = \frac{\mathbb{E}[\sum_i X_i]}{n} = \frac{n\mathbb{E}[X_i]}{n} = \mathbb{E}[X_i]$$

So, convergence in expectation is even weaker than WLLN (convergence in probability).

We will show the first implication by showing something stronger; that almost sure convergence implies convergence in probability.

Proof 3.1 (A.S. implies Convergence in Probability) Fix $\varepsilon > 0$ and let X_n converge to X almost surely. Define the following events:

$$\begin{aligned} A_n &= \{|X_n - X| > \varepsilon\} \\ B_n &= \cup_{m\geq n} A_m \\ B &= \cap_{n=1}^{\infty} B_n \\ B &= \{A_n \text{i.o.}\} \end{aligned}$$

However, by definition of almost sure convergence, we must have $\mathbb{P}[A_n \text{i.o.}] = 0$. However:

$$\begin{aligned} \mathbb{P}[|X_n - X| > \varepsilon] &\leq \mathbb{P}[B_n] \to 0 \\ \mathbb{P}[|X_n - X| > \varepsilon] &\to 0 \end{aligned}$$

However, convergence in probability does not always imply almost sure convergence. Here is a counter-example:

Example 3.2 (Not all convergence is made equal) Pick ω uniformly in $[0, 1]$.
Define $X_1(\omega) = 1$ and then for $n \geq 2$, make it piecewise, with value 1 for a width $1/n$ and value 0 for a width $1 - 1/n$. Furthermore, we stack the start of the pulse at the end of the interval of the last one; wrapping it around as needed (modulo 1).
For any ε, the width of X_n being 1 goes down to 0, as we continue along.

$$\mathbb{P}[|X_n - 0| > \varepsilon] \to 1$$

However,

$$\mathbb{P}\left[X_n \to 0\right] = 0$$

since for some ω, we can always find an X_n later which is 1 (there will be another pulse containing ω).

Armed with the machinery, now we can finally prove the Big Theorem.

Proof 3.2 Let us focus on part a). We define:

$$m_j = \mathbb{E}\left[T_j \mid X(0) = j\right]$$

where $T_j = \min\{n > 0 \mid X(n) = j\}$ (i.e. first time visiting j).
Then, we claim the following equality holds (for large N).

$$N \sum_j \frac{1}{m_j} P(j,i) = \frac{N}{m_i}$$

Each term in the left hand side is the long term fraction of time you've spent in visiting j and then i, while the term on the right hand side is the long term fraction of time you've spent visiting i.
Dividing through by N, we have that:

$$\sum_j \frac{1}{m_j} P(j,i) = \frac{1}{m_i}$$

calling $\pi = \begin{bmatrix} \frac{1}{m_1} & \frac{1}{m_2} & \dots \end{bmatrix}$, then we essentially have:

$$\pi(j)P(j,i) = \pi(i)$$
$$\pi P = \pi$$

Proving uniqueness is simple and was left out of lecture.

Next let us prove part b). Let $A(n)$ be the number of visits to state i by time n. Let T_j^ℓ be the difference between ℓth time you came back to j and the $\ell - 1$th time. Then:

$$A(n)/n \sim \frac{k}{T_i^1 + T_i^2 + \cdots + T_i^k}$$

since the LHS is frequency. However, note that T_i^ℓ are i.i.d. because the Markov chain is memoryless, so coming back to yourself the next time Then:

$$\frac{1}{n/A(n)} = \frac{T_i^1 + T_i^2 + \cdots + T_i^k}{k}$$

However, the RHS almost surely goes to the average value, m_i. This means

$$\frac{1}{n/A(n)} \to \frac{1}{m_i} = \pi(i)$$

4 Multiplexing

4.1 Lecture 9, Continued

Multiplexing relates to the sharing of a common resource. Consider some link or channel with some rate capacity C. Suppose it is being shared by some group of connections $x_1, x_2, \ldots, x_n$, but not all connections are active at the same time. Think about a phone line–everyone doesn't have to use the phone at the same time. Then, the rate for each connection is C/v, where v is a random variable the represents the number of active connections.

Let us model link sharing with each channel being active as a Bernoulli random variable with probability p of the link being active. Then,

$$v \sim B(n, p)$$

We also discuss the following:

Definition 4.1 (PPF) The percent point function (PPF) of a random variable is the "inverse" CDF. Note that a lot of CDFs are not bijections, so instead we say the following.
Suppose $F_X(x) = p_1$ for all $x \in [x_1, x_2)$ and then it jumps to p_2 at x_2, the PPF as follows:

$$PPF(p) = x_2 \qquad \forall p \in (p_1, p_2]$$
$$PPF(p) = x_1 \qquad p = p_1$$

Then suppose we wanted to find the smallest m such that $\mathbb{P}[v > m] \leq \delta$ or $\mathbb{P}[v \leq m] \geq 1 - \delta$. If δ is small, this means that each active user will get at least a rate of C/m with probability $1 - \delta$ or higher (think of this rate as the speed of your internet connection, for example). To do this, just take $PPF(1 - \delta)$ to get the correct m to be confident (typically $\delta = 0.05$).

Now, we return to Normal/Gaussian random variables, as they help us investigate the sums of random variables. Here's a few useful facts:

Note 4.1 For $X \sim \mathcal{N}(\mu, \sigma^2)$:

- $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$, $-\infty < x < \infty$.
- $X = \mu + \sigma W$, where $W \sim \mathcal{N}(0, 1)$ (the standard normal). You can show this by change of variables on the CDF.
- $\mathbb{P}[W > 1.65] \approx 0.05$, $\mathbb{P}[W > 1.96] \approx 0.025$, $\mathbb{P}[W > 2.32] \approx 0.01$
- The above facts holds for general Gaussian RV (i.e. if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{P}[X > \mu + 1.65\sigma] = 0.05$)
- The normal is symmetric about its mean.

4.2 Lecture 10

4.2.1 Central Limit Theorem

First, we talk about another type of convergence.

Definition 4.2 (Convergence in Distribution) Let $\{X(n), n \geq 1\}$ and X be random variables. We say $X(n)$ converges in distribution to X (or it weakly converges), and write $X(n) \xrightarrow{d} X$, if

$$\mathbb{P}[X(n) \leq x] \to \mathbb{P}[X \leq x] \; \forall x \text{ s.t. } \mathbb{P}[X = x] = 0$$

Let us see why the $\mathbb{P}[X = x] = 0$ is necessary.

Example 4.1 Consider $X(n) = 3 + \frac{1}{n}$ and $X = 3$. We want these to converge in distribution, but consider that

$$\mathbb{P}[X(n) \leq 3] = 0$$

for all n, but $\mathbb{P}[X \leq 3] = 1$, which would lead a lack of convergence. However, we stipulate that we should only consider points x where there is no discrete mass; $x = 3$ does not qualify because $\mathbb{P}[X = 3] = 1 > 0$. So, the convergence in distribution still happens!

This is the weakest type of convergence we have discussed. If we know convergence in probability of $X(n)$ to X, then we can conclude that $X(n) \xrightarrow{d} X$. Let us see a counter example of the converse.

Example 4.2 Take $X_n = X \sim \text{Bernoulli}(1/2)$. Note that $X_n \xrightarrow{\text{d}} 1 - X$ because the CDFs are identical. However, take $0 < \varepsilon < 1$. Then:

$$\begin{aligned}\mathbb{P}[|X_n - (1 - X)| > \varepsilon] &= \mathbb{P}[|X - (1 - X)| > \varepsilon] \\ &= \mathbb{P}[|2X - 1| > \varepsilon] \\ &\to 1\end{aligned}$$

We now have an important result, armed with our Gaussian knowledge.

Theorem 4.1 (Central Limit Theorem) Let $\{X(n), n \geq 1\}$ be a set of iid random variable with mean $\mathbb{E}[X(n)] = \mu$ and variance $\text{Var}(X(n)) = \sigma^2$. Define $S(n) = \sum_{i=1}^{n} X(i)$. Then,

$$\frac{S(n) - n\mu}{\sigma\sqrt{n}} \xrightarrow{\text{d}} \mathcal{N}(0, 1)$$

In practice, we can think of CLT in another way. Let X_i be iid random variables with mean μ and variance σ^2. Then

$$X_1 + X_2 + \cdots + X_n \approx \mathcal{N}\left(n\mu, n\sigma^2\right)$$

Here is a quick result, based on the fact that a Binomial RV is the sum of many Bernoulli RVs.

Theorem 4.2 If $X_N \sim \text{Binomial}(N, p)$, then:

$$X_N \approx \mathcal{N}(Np, Np(1 - p))$$

for large N.

Now let us see how to apply CLT to find bounds. Let $Y(N) \sim \text{Binomial}(N, p)/N$. Then

$$\sigma_{Y(N)} = \sqrt{p(1-p)N/N^2} = \sqrt{p(1-p)/N}$$

and the mean of $Y(N)$ is p. Define

$$A_1 = \{Y(N) \geq p + 1.65\sqrt{p(1-p)/N}\}$$
$$A_2 = \{Y(N) \leq p - 1.65\sqrt{p(1-p)/N}\}$$

Due to CLT, $\mathbb{P}[A_1 \cup A_2] \approx 0.1$ (from the tail identities we covered last lecture) and thus $\mathbb{P}\left[A_1^C \cap A_2^C\right] \approx 0.9$, i.e.

$$\mathbb{P}\left[p - 1.65\sqrt{p(1-p)/N} \leq Y(N) \leq p + 1.65\sqrt{p(1-p)/N}\right] = 0.9$$
$$\mathbb{P}\left[Y(N) - 1.65\sqrt{p(1-p)/N} \leq p \leq Y(N) + 1.65\sqrt{p(1-p)/N}\right] = 0.9$$

This statement states that $Y(N)$ as an estimator for p is within this interval with confidence 90Replacing the argument with 1.96 gives 95% confidence. Sometimes, we argue that $p(1-p) < 1/4$, so

$$\mathbb{P}\left[Y(N) - 1.65\sqrt{0.25/N} \leq p \leq Y(N) - 1.65\sqrt{0.25/N}\right] = 0.9$$
$$\mathbb{P}\left[Y(N) - \frac{0.83}{\sqrt{N}} \leq p \leq Y(N) - \frac{0.83}{\sqrt{N}}\right] = 0.9$$

However, when we only have observations and do not know the underlying distribution, we can do the following.

Note 4.2 Consider any iid random variables $\{X(n), n \geq 1\}$ without knowledge of variance.

$$\mu_n = \frac{X(1) + X(2) + \cdots + X(n)}{n}$$
$$\sigma_n^2 = \frac{\sum_{m=1}^{n}(X(m) - \mu_n)^2}{n-1}$$

where the $n-1$ comes from the fact that we want σ_n^2 in expectation to be the same as the true variance of the $X(i)$'s. (see Walrand for derivation).
Then, the following is a 90% confidence interval for $\mu = \mathbb{E}[X(i)]$:

$$[\mu_n - 1.65\frac{\sigma_n}{\sqrt{n}} < \mu < \mu_n + 1.65\frac{\sigma_n}{\sqrt{n}}]$$

and this is a 95% confidence interval for μ:

$$[\mu_n - 2\frac{\sigma_n}{\sqrt{n}} < \mu < \mu_n + 2\frac{\sigma_n}{\sqrt{n}}]$$

4.3 Lecture 11

4.3.1 Characteristic and Moment-Generating Functions

Definition 4.3 (Characteristic Function) The characteristic function of a random variable X is the function:

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

with domain $u \in \mathbb{R}$ and $i = \sqrt{-1}$.

Definition 4.4 (Moment-Generating Function) The moment-generating function (MGF) of a random variable X is the function:

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

with domain $t \in \mathbb{R}$.

The characteristic function and moment generating function uniquely determines a random variable. For example, you can determine the associated PDF/CDF from it.

Note that the MGF does not always exist (if X is large with high probability, it may blow up), but when it does the relationship between it and the characteristic function can be summarized as:

$$\phi_X(t) = M_{iX}(t) = M_X(it)$$

Furthermore, we have the following:

Note 4.3 (Moment Generation)

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

So,

$$\mathbb{E}\left[X^n\right] = M_X^{(n)}(0)$$

i.e. the nth derivative of M_X evaluated at 0.

We work through an example.

Example 4.3 (Characteristic of $\mathcal{N}(0,1)$) We apply the definition with LOTUS directly.

$$\begin{aligned} \phi_X(u) &= \int_{-\infty}^{\infty} e^{iux} \frac{1}{2\pi} e^{-x^2/2} \, \mathrm{d}x \\ \frac{\mathrm{d}}{\mathrm{d}u}\phi_X(u) &= \int_{-\infty}^{\infty} ixe^{iux} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, \mathrm{d}x \\ &= -\int_{-\infty}^{\infty} ie^{iux} \frac{1}{2\sqrt{\pi}} \, \mathrm{d}e^{-x^2/2} \end{aligned}$$

Now we apply integration by parts:

$$\begin{aligned}
\phi_X'(u) &= -ie^{iux}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} i\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,\mathrm{d}e^{iux} \\
\phi_X'(u) &= -u\int_{\infty}^{\infty} e^{iux}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,\mathrm{d}x \\
\phi_X'(u) &= -u\phi_X(u) \\
\frac{\phi_X'(u)}{\phi_X(u)} &= -u \\
(\log(\phi_X(u)))' &= -u \\
\log(\phi_X(u)) &= -u^2/2 + c \\
\phi_X(u) &= Ae^{-u^2/2}
\end{aligned}$$

Furthermore, we know that $\phi_X(0) = 1$ (since this is just the integral of the Normal over $\mathbb{R}$). Thus,

$$\phi_X(u) = e^{-u^2/2}$$

Now, let us use this form to find the moments of $\mathcal{N}(0,1)$.

$$\begin{aligned}
\phi_X(u) &= \mathbb{E}\left[e^{iuX}\right] \\
&= \sum_{n=0}^{\infty}\frac{1}{n!}i^n u^n \mathbb{E}\left[X^n\right] \\
&= e^{-u^2/2} \\
&= \sum_{m=0}^{\infty}\frac{1}{m!}\left(\frac{-u^2}{2}\right)^m
\end{aligned}$$

Now, we match coefficients of u^{2m}.

$$\begin{aligned}
\frac{1}{(2m)!}i^{2m}\mathbb{E}\left[X^{2m}\right] &= \frac{1}{m!}\left(\frac{-1}{2}\right)^m \\
\mathbb{E}\left[X^{2m}\right] &= \frac{(2m)!}{m!\cdot 2^m}
\end{aligned}$$

Note that the summation has no terms for u^{2m+1}, so the coefficients are all zero. This means $\mathbb{E}\left[X^{2m+1}\right] = 0$. Altogether:

$$\mathbb{E}\left[X^n\right] = \begin{cases} \frac{n!}{(n/2)!\cdot 2^n} & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd} \end{cases}$$

Here is a useful formula for MGFs:

Theorem 4.3 (Independent MGFs) Consider indepdent random variables X_1 and X_2. Then the moment-generating function of their sum is given by:

$$M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t)$$

4.3.2 Proof of the CLT

We give a brief sketch of the proof of the Central Limit Theorem. This is not a full argument, as we do not show that the characteristic function uniquely determines a random variable.

Proof 4.1 Define

$$Y(n) = \frac{X(1) + X(2) + \cdots + X(n) - n\mu}{\sigma\sqrt{n}}$$

Then, let us compute its characteristic function.

$$\begin{aligned}
\phi_{Y(n)}(u) &= \mathbb{E}\left[e^{iuY(n)}\right] \\
&= \mathbb{E}\left[\prod_{m=1}^{n} \exp\left\{\frac{iu(X(m)-\mu)}{\sigma\sqrt{n}}\right\}\right] \\
&= \prod_{m=1}^{n} \mathbb{E}\left[\exp\left\{\frac{iu(X(m)-\mu)}{\sigma\sqrt{n}}\right\}\right] && \text{(Independence)} \\
&= \mathbb{E}\left[\exp\left\{\frac{iu(X(1)-\mu)}{\sigma\sqrt{n}}\right\}\right]^n && \text{(Identically Distributed)} \\
&= \mathbb{E}\left[1 + \frac{iu(X(1)-\mu)}{\sigma\sqrt{n}} - \frac{u^2(X(1)-\mu)^2}{2\sigma^2 n} + o\left(\frac{1}{n}\right)\right]^n \\
&= \left(1 - \frac{iu(\mathbb{E}\left[X(1)-\mu\right])}{\sigma\sqrt{n}} - \frac{u^2}{2}\cdot\frac{(X(1)-\mu)}{\sigma^2} + o\left(\frac{1}{n}\right)\right)^n \\
&= \left(1 - \frac{u^2}{2} + o\left(\frac{1}{n}\right)\right)^n \\
&\to e^{-u^2/2} && \text{(Limit Definition of } e\text{)}
\end{aligned}$$

Since the characteristic function converges to that of a standard normal, this means that the distribution converges as well.

4.4 Lecture 12

4.4.1 Limits of Distributions

Now we discuss results which are applications of Characteristic Functions.

Theorem 4.4 Consider fixed $\lambda > 0$. Then we have,

$$\lim_{n\to\infty} B(n, \lambda/n) = P(\lambda)$$

Proof Although this is easy to see by taking the limit of PDFs, we opt for characteristic functions instead. Let

$$X_n \equiv_D B(n, \lambda/n)$$

Then,

$$\begin{aligned}\phi_{X_n}(u) &= \mathbb{E}\left[e^{iuX_n}\right] \\ &= \mathbb{E}\left[e^{iu(Z_n(1)+\cdots+Z_n(n))}\right]\end{aligned}$$

where $Z_n(i)$ are iid $B(\lambda/n)$.

$$\begin{aligned}\phi_{X_n}(u) &= \prod_{i=1}^{n} \mathbb{E}\left[e^{iuZ_n(i)}\right] \\ &= \left(\frac{\lambda}{n}e^{iu0} + \left(1-\frac{\lambda}{n}\right)e^{iu1}\right)^n \\ &= \left(1+\frac{\lambda}{n}\left(e^{iu}-1\right)\right)^n \\ &\to_{n\to\infty} \exp\left(\lambda(e^{iu}-1)\right)\end{aligned}$$

However, if $X \equiv_D P(\lambda)$, then its characteristic function is:

$$\begin{aligned}\phi_X(u) &= \sum_{m=0}^{\infty} \frac{\lambda^m e^{-\lambda}}{m!} e^{ium} \\ \phi_X(u) &= e^{-\lambda} \sum_{m=0}^{\infty} \frac{\left(\lambda e^{iu}\right)^m}{m!} \\ \phi_X(u) &= e^{-\lambda} \exp\left(\lambda e^{iu}\right) \\ \phi_X(u) &= \exp\left(\lambda\left(e^{iu}-1\right)\right)\end{aligned}$$

So the characteristic function $\phi_{X_n}(u)$ converges to $\phi_X(u)$, so there is a convergence in distribution.

The next result is similar, and can be proved in the same way (proof omitted here).

Theorem 4.5 Consider fixed $\lambda > 0$. Then we have,

$$\lim_{n\to\infty} \frac{G\left(\frac{\lambda}{n}\right)}{n} = Exp(\lambda)$$

Here are some more results of manipulation:

Theorem 4.6 Let $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$ be independent. Then: $X^2 + Y^2 \sim \text{Exp}(1/2)$.

Proof Let $Z = X^2 + Y^2$. Then:

$$\begin{aligned}
\mathbb{P}[Z \le z] &= \mathbb{P}\left[X^2 + Y^2 \le z\right] \\
&= \frac{1}{2\pi} \int \int_{X^2+Y^2 \le z} e^{-(x^2+y^2)/2} \, \mathrm{d}x \, \mathrm{d}y \\
&= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\sqrt{z}} e^{-r^2/2} r \, \mathrm{d}r \, \mathrm{d}\theta \\
&= \int_0^{\sqrt{z}} r e^{-r^2/2} \, \mathrm{d}r \\
&= -e^{-r^2/2} \Big|_0^{\sqrt{z}} \\
&= 1 - e^{-z/2}
\end{aligned}$$

Which exactly the CDF of an exponential.

Theorem 4.7 Consider $X \sim \mathcal{N}(0,1)$. Let $Q(x) = P(X > x)$ (called the error function). Then:

$$\frac{x}{1+x^2} f_X(x) \le Q(x) \le \frac{1}{x} f_X(x)$$

Now we return to the application of multiplexing. If there is a lot of traffic at an output port of a switch, this can cause buffering (building a buffer of things to transmit next, if there's too many to transfer all at once). How can we model buffering? Let's try using a discrete-time Markov chain.

Suppose that at each instant n^+, a packet arrives with probability $\lambda \in [0,1]$ independently of the past. The time to transmit a packet is geometrically distributed with parameter $\mu \in (0,1]$. All transmission times are independent. A packet in service completes transmission at n^- with probability μ.

However, for the purposes of this course, this DTMC model is too complicated. We instead use CTMCs (Continuous-Time Markov Chains).

5 Networks

5.1 Lecture 12, Continued

We start our discussion of networks with the familiar Markov Chains, but with a twist.

Definition 5.1 (Infinite DTMC) You can consider $\{X(n), n \ge 0\}$ as a Markov Chain over a (countably) infinite state space $\mathcal{X} = \{0, 1, 2, \dots\}$.
The initial distribution is $\pi(i), i \in \mathcal{X}$ such taht $\pi(i) \ge 0$ and $\sigma_i \pi(i) = 1$. The state transition probability matrix P also exists, where $\sum_j P(i,j) = 1$.

Irreducibility and aperiodicity are defined the same way as in the finite case. The balance equations also hold for the invariant distribution, as $\pi = \pi P$.

Definition 5.2 (Types of States) A state is **transient** if starting the Markov Chain from this state causes it to be visited finitely often (i.e. there is some probability > 0 that you'll never come back to it). A state is **recurrent**

if it's not transient.
A recurrent state is **positive recurrent** if the average time between successive visits is finite, otherwise it's **null recurrent**.

Transience and recurrence make sense in finite DTMCs, but positive and null recurrence only make sense in infinite DTMCs.

Theorem 5.1 For an irreducible DTMC, states are either all transient, all positive recurrent, or all null recurrent.

There is also the idea of communicating classes. Basically, these are strongly connected components in the digraph representing the Markov chain. A communicating class is a maximal set of states such that there is a nonzero probability path from any one state to any other.

A good example of an infinite DTMC is the random walk reflected at 0, i.e. at any natural number you can walk right or left with probabilities p and $1-p$ respectively; at 0 you walk to the right with probability p and stay in 0 with probability $1-p$.

Then, we have the following:

- If $p > 1/2$, then it is a transient MC.
- If $p = 1/2$, then it is a null-recurrent.
- If $p < 1/2$, then it is a positive-recurrent.

Proof We talk about the three cases separately, setting $p + q = 1$:

- $p > q (p > 1/2)$ Let $Z(n)$ be iid with $\mathbb{P}[Z(n) = 1] = p, \mathbb{P}[Z(n) = -1] = q$. Now define:

$$\begin{aligned} X(n) &= \max(X(n-1) + Z(n), 0) \\ X(n) &\geq X(0) + Z(1) + \cdots + Z(n) \\ \frac{X(n)}{n} &\geq \frac{X(0) + Z(1) + \cdots + Z(n)}{n} \\ &\to \mathbb{E}[Z] \end{aligned}$$

  However, note that $\mathbb{E}[Z] = 1 \cdot p - 1 \cdot q > 0$, so $X(n) = \lim_{n\to\infty} n(p-q) \to \infty$, meaning the Markov chain is transient.

- $p < q (p < 1/2)$ Note that $p\pi(n) = q\pi(n+1)$ by the balance equations. Define $\rho = \frac{p}{q}$. This means

$$\pi(n+1) = \frac{p}{q}\pi(n) \implies \pi(n)(1-\rho)\rho^n$$

  This means that $\pi(i) = \frac{1}{\mathbb{E}[T_i | X(0)=i]}$ meaning the average time to return to state i is finite (since $\frac{1}{\pi(i)}$ is finite). Thus, this Markov chain is positive-recurrent.

- $p = q (p = 1/2)$ We investigate $\mathbb{E}[T_0 \mid X(0) = 1]$. By means of first-step equations, this is:

$$\mathbb{E}[T_0 \mid X(0) = 1] = 1 + \frac{1}{2}\mathbb{E}[T_0 \mid X(0) = 2]$$

  By symmetry,

$$\mathbb{E}[T_0 \mid X(0) = 2] = 2\mathbb{E}[T_0 \mid X(0) = 1]$$

  However, this means:

$$\begin{aligned} \mathbb{E}[T_0 \mid X(0) = 1] &= 1 + \mathbb{E}[T_0 \mid X(0) + 1] \\ 1 &= 0 \end{aligned}$$

Thus, the expectation must be infinite! This means that:

$$\mathbb{E}[T_0 \mid X(0) = 0] = 1 + \frac{1}{2}\mathbb{E}[T_0 \mid X(0) = 1] = \infty$$

Now, we just need show that the probability of returning to a state is 1. Here is a sketch of the argument. Suppose we are trying to return to state n. If we go to $n-1$, then eventually we will come back to n, this is intuitive (think about the ants homework problem). If we go to $n+1$, then to return, we have to go to the left a net amount of 1; i.e. it's like flipping a bunch of fair coins and having at some point more tails then heads. However, this happens with probability 1 (which can be proved with Borel-Cantelli).

Thus, this Markov chain is null-recurrent.

We also have to modify the big theorem for infinite DTMCs.

Theorem 5.2 Consider an irreducible DTMC $\{X(n), n \geq 0\}$ over an infinite state space with an invariant distribution π. Then for each $i, \pi(i) = \frac{1}{\mathbb{E}[T_i | X(0)=i]}$ where T_i is the first time after 0 to reach state i. Furthermore,

1. If the markov chain is positive recurrent, there is a unique invariant distribution π where

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{1}_{\{X(n)=i\}} = \pi(i)$$

   i.e. the long-term fraction of time converges to the invariant distribution.

2. If the markov chain is positive recurrent and aperiodic, $\pi_n \to \pi$ regardless of the initial distribution.

3. If the markov chain is not positive recurrent, it does not have an invariant distribution, and the fraction of time spent in any one state goes to 0.

We again try to apply to balance equations to our random walk reflected at 0. However, we can consider many states at once as a kind of "meta-state." Then, we track only the "net" flow in and out of our "meta-state." These must be equal in the balance equations (think of the internal transitions as internal forces in mechanics; by Newton's second law, these cancel out).

Let us make our meta-state $\{0, 1, \ldots, n\}$. The flow out is $p\pi(n)$ and the flow in is $(1-p)\pi(n+1)$. Calling $\rho = \frac{p}{1-p}$. This means:

$$\begin{aligned} p\pi(n) &= (1-p)\pi(n+1) \\ \pi(n) &= \rho^n \pi(0) \\ \sum_{i=0}^{\infty} \rho^i \pi(0) &= 1 \\ \pi(0) &= \frac{1}{1-\rho} \end{aligned}$$

Note that if $p = 1/2$, then $\rho = 1$.

Now, we define another condition:

Definition 5.3 (Detailed Balance Equations) The condition $\pi(i)P(i,j) = \pi(j)P(j,i)$ for any two states i and j are called detailed balance equations.

This notion will be important later (when we discuss reversibility).

5.2 Lecture 13

5.2.1 Poisson Process

Definition 5.4 (Poisson Process) Let $\lambda > 0$ and $\{S_1, S_2, \dots\}$ be IID $Exp(\lambda)$ random variables, i.e. inter-arrival times. Also, let $T_n = S_1 + \cdots + S_n$ for $n \geq 1$, i.e. the nth arrival time. Define $N_t = 0$ if $t < T_1$, otherwise $N_t = \max\{n \geq 1 \mid T_n \leq t\}, t \geq 0$. Then, $N = \{N_t, t \geq 0\}$ is a Poisson process with rate λ.

Intuitively, N_t counts number of jumps (or arrivals) that have occurred by time t; every Expo(λ) arrival time (S_i), there is a new arrival. The arrival times are iid.

Theorem 5.3 (Memorylessness of the Poisson Process) Let $N = \{N_t, t \geq 0\}$ be a Poisson process with rate λ. Given $\{N_s, s \leq t\}$, then $\{N_{s+t} - N_t, s \geq 0\}$ is a Poisson process with rate λ.

The reason for this is because of the memorylessness of the exponential random variable! After time t, the next "jump" is still Expo(λ), because it is basically $\{T_m - t \mid T_m \geq t\}$.

Theorem 5.4 Any Poisson process has independent and stationary increments. To be precise, for any $0 \leq t_1 < t_2 < t_3 < \dots$ increments, $\{N_{t_{n+1}} - N_{t_n}\}$ are independent for all n and the distribution of $N_{t_{n+1}} - N_{t_n}$ depends solely on $t_{n+1} - t_n$ (it is 'stationary', i.e. absolute time-invariant).

In fact, the distribution of N_t is easy to find.

Theorem 5.5 (Poisson Process is a Poisson Distribution) Let $N = \{N_t, t \geq 0\}$ be a Poisson process with rate λ. Then $N_t \sim$ Poisson (λt).

Proof Let T_n be the nth jump time and let $S_n = T_n - T_{n-1}$, $n \geq 1$, $T_0 = 0$, i.e. S_n's are the inter-arrival times. Then, we can compute the probability of having exactly n arrivals in the interval 0 to t around specific times $0 \leq t_1 < t_2 < \cdots < t_n$

$$
\begin{aligned}
&= \mathbb{P}\left[T \in (t_1, t_1 + \mathrm{d}t_1), T_2 \in (t_2, t_2 + \mathrm{d}t_2), \dots, T_n \in (t_n, t_n + \mathrm{d}t_n), T_{n+1} > t\right] \\
&= \mathbb{P}\left[S_1 \in (t_1, t_1 + \mathrm{d}t_1), \dots, S_n \in (t_n - t_{n-1}, t_n - t_{n-1} + \mathrm{d}t_n), S_{n+1} > t - t_n)\right] \\
&= \mathbb{P}\left[S_{n+1} > t - t_n\right] \Pi_{i=1}^{n} \lambda e^{-\lambda(t_i - t_{i-1})}\, \mathrm{d}t_i \\
&= (e^{-\lambda(t-t_n)}) \Pi_{i=1}^{n} \lambda e^{-\lambda(t_i - t_{i-1})}\, \mathrm{d}t_i \\
&= \lambda^n e^{-\lambda t} \Pi_{i=1}^{n}\, \mathrm{d}t_i
\end{aligned}
$$

Note that this does not depend on the actual t_i's, only the difference between them. Thus, any t_i's could have been chosen and been equally likely. Thus, have exactly n arrivals in the interval 0 to t aroudn any times is just the integral over the volume of all the possibilities $S = \{(t_1, t_2, \dots, t_n) \mid \text{all } t_i\text{'s strictly increasing}\}$. Note that the volume of $|S| = \frac{t^n}{n!}$ because normally we would have a volume of t^n, but only one ordering of these is strictly increasing. Thus, we take the integral to use total probability rule.

$$
\begin{aligned}
\mathbb{P}\left[N_t = n\right] &= \int_S \lambda^n e^{-\lambda t} \Pi_{i=1}^{n}\, \mathrm{d}t_i \\
&= \lambda^n e^{-\lambda t} \cdot |S| \\
&= \frac{(\lambda t)^n e^{-\lambda t}}{n!}
\end{aligned}
$$

Which is precisely the PMF of a Poisson distribution.

Here is another way to think of Poisson Processes; you can simulate them two ways:

- Generate N with a Posson distribution with parameter λt.
- If $N = n \geq 1$, generate $U_1, U_2, \ldots, U_n$ iid distribution Uniform$[0, 1]$ and order them (order statistics) to get, $U_{(1)}, U_{(2)}, \ldots, U_{(n)}$. Then, set the arrival times to be:

$$(T_1, \ldots, T_n) = (U_{(1)}, \ldots, U_{(n)})$$

The second way comes from the fact that the ordered arrival times between 0 to t have uniform density $f(t_1, t_2, \ldots, t_n \mid N_t = n) = \frac{n!}{t^n}$ (thus unordered ones are uniform with probability $\frac{1}{t^n}$).

5.2.2 Continuous Time Markov Chains

We begin again with the definition.

Definition 5.5 (Countinuous Time Markov Chain) Let $\mathcal{X}$ be a finite or countable state space, and define a rate matrix $Q = \{q(i, j), i, j \in \mathcal{X}\}$ such that $q(i, j) \geq 0$ for all $i \neq j$ and $\sum_j q(i, j) = 0$ for all i (this means the diagonal entries in each row are the negative sum of all the other entries in that row).
A continuous-time Markov chain (CTMS) with initial distribution and rate matrix Q is a process $\{X_t, t \geq 0\}$ such that $\mathbb{P}[X_0 = i] = \pi_0(i)$ and the transition property:

$$\mathbb{P}\left[X_{t+\epsilon} \mid X_t = i, X_u, u < t\right] = \mathbb{1}_{\{i=j\}} + \epsilon q(i, j) + o(\epsilon)$$

Another way to consider this construction is as follows: define $q_i = -q(i, i)$. We stay in some state i until some time $\tau \sim \text{Exp}(q_i)$. If $X_t = i$, then $X_{t+\tau} = j$ (it jumps to $j \neq i$) with probability $\Gamma(i, j) = \frac{q(i,j)}{q_i}$. Note that:

$$\sum_{j \neq i} \Gamma(i, j) = 1$$

Note that this is equivalent to taking the following: when you reach state i, start a clock for each of the other states which is exponential with parameter $q(i, j)$. Whichever clock jumps first, we go to that state. The reason this is equivalent is because this is taking the min of all these timers, which creates an exponential with parameter $\sum_{i \neq j} q(i, j) = q_i$, as we have in the original construction. Furthermore, we know

$$\mathbb{P}\left[\min_{j \neq i} T_j = T_k\right] = \frac{q(i, k)}{\sum_{j \neq i} q(i, j)} = \Gamma(i, j)$$

5.3 Lecture 14

Most of the lecture was going through some claims in previous lectures; I updated those sections accordingly.

Definition 5.6 (Stopping Time) Stopping time is a random variable τ that has the property $\{\tau \leq t\}$ is completely determined by the process $\{X_s, s \leq t\}$.

An example of a stopping time is as follows.

Example 5.1 (Gambling) Suppose we start with \$10 and bet \$1 each game, where we win \$1 with probability $p < 1/2$ and lose our money otherwise.
The time to go bust is a stopping time, since it does not depend on knowing the evolution of the Markov chain after that time.
The last time to have \$5 is not a stopping time, since it depends on knowing when you go bust (the future of that stopping time).

This leads to the following property (called the "Strong Markov Property").

Theorem 5.6 (Strong Markov Property) If τ is a stopping time,

$$\mathbb{P}\left[\{X_{\tau+t}, t \geq 0\} \in A \mid X_\tau = k, X_t, t \leq \tau\right] = \mathbb{P}\left[\{X_t, t \geq 0\} \in A \mid X_0 = k\right]$$

Note that this is stronger than the normal Markov property (τ can be an RV instead of a constant).

For CTMCs and DTMCs that we discuss, the Strong Markov Property holds.

5.4 Lecture 15

We next discuss the invariant distribution of a CTMC.

Theorem 5.7 (Invariant Distribution of a CTMC) Let π_t be the distribution of X_t.
Note that

$$\pi_{t+\epsilon}(i) \approx \sum_{j \neq i} \pi_t(j) q(j,i)\epsilon + \pi_t(i)(1 - q_i\epsilon)$$

which can be rewritten as: $\pi_{t+\epsilon} = \pi_t(I + Q\epsilon)$, or $\frac{\pi_{t+\epsilon} - \pi t}{\epsilon} = \pi_t Q$
Hence by limits, $\frac{\mathrm{d}}{\mathrm{d}t}\pi_i = \pi_t Q$, which we call the Kolmogorov Forward Equation. This means that:

$$\pi_t = \pi_0 e^{Qt} = \pi_0\left(I + Qt + \frac{1}{2!}Q^2t^2 + \frac{1}{3!}Q^3t^3 + \dots\right)$$

Now we will show $\pi_t = \pi, t \geq 0 \iff \pi_0 = \pi, \pi Q = 0$. The forward direction is:

$$\pi_0 = \pi$$
$$\frac{\mathrm{d}\pi_t}{\mathrm{d}t} = 0 = \pi_t Q = \pi Q$$

Where the derivative equation
The backwards direction:

$$\pi_t = \pi_0 e^{Qt}$$
$$\pi_t = \pi(I + Qt + \dots)$$
$$\pi_t = \pi$$

as required.
Thus, to find the invariant distribution, we must have $\pi Q = 0$ and $\sum \pi(i) = 1$.

To see why the former condition is "flow-in = flow-out", let us carry out the matrix vector multiplication element-wise:

$$\pi Q = 0$$
$$\sum_i \pi(i) q(i,k) = 0$$
$$\sum_{i \neq k} \pi(i) q(i,k) + \pi(k) q(k,k) = 0$$
$$\sum_{i \neq k} \pi(i) q(i,k) = \sum_{j \neq k} \pi(k) q(k,j)$$

The LHS is the flow into state k; the RHS is flow out of state k.

We also have some approximative notions.

5.4.1 DTMC Approximation of CTMC

Definition 5.7 (Embedded DTMC) The embedded DTMC (or Jump DTMC) of a CTMCs has probability transition matrix $P = \Gamma$ where

$$\Gamma(i,j) = \begin{cases} \frac{q(i,j)}{q_i} & i \neq j \\ 0 & i = j \end{cases}$$

Theorem 5.8 If the invariant distribution of the Jump DTMC satisfies $\nu = \nu\Gamma$, we have the following relationships between ν and π (the invariant distribution of the original CTMC):

$$\pi(i) = \frac{\nu(i)/q_i}{\sum_k \nu(k)/q_k}$$
$$\nu(i) = \frac{\pi(i)q_i}{\sum_k \pi(k)q_k}$$

Proof We give an intuitive argument. Note that $\nu(i) \cdot N$ gives the amount of epochs spent in some state i in the DTMC and $\frac{1}{q_i}$ is the average amount of time spent in state i when you're there: the long term (continuous) time spent in state i is $\frac{\nu(i)}{q_i}$. Thus, this means that the average amount of (continuous) time in state i, $\pi(i)$ is:

$$\pi(i) = \frac{\nu(i)/q_i}{\sum_k \nu(k)/q_k}$$

For $\nu(i)$, we have similar reasoning. Since each epoch lasts on average $\frac{1}{q_i}$, then $\frac{\pi(i)}{\frac{1}{q_i}}$ gives you the amount of times we visited state i when running the Markov chain. Thus, the ratio of the fraction of visits to total visits is $\nu(i)$, which is consistent with the above formulation.

Although these probabilities wouldn't capture the time spent "sojourning," it is still a decent approximation.

Definition 5.8 (Approximate DTMC) We instead consider every time step in some CTMC to have size ϵ. If $q(i, j) = \lambda$, then $P(i, j) = \lambda\epsilon$. This yields a DTMC with $P = I + Q\epsilon$.
Note that the invariant distribution is the same as that of the CTMC.

We now extend our usual Markov chain theorems to CTMC.

Theorem 5.9 For an irreducible CTMC, states are either all transient, all positive recurrent, or all null recurrent.

Theorem 5.10 (Big Theorem) Consider an irreducible CTMC over a countable state space. Then,

- If it's positive recurrent, there is a unique invariant distribution π.
- If it's positive recurrent, long-term fraction of time converges to the invariant distribution:

$$(X_t = i) = \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbb{1}_{X_s=i} \, \mathrm{d}s = \pi(i)$$

- If it's positive recurrent, $\pi_t \to \pi$.
- If not positive recurrent, it does not have an invariant distribution, and fraction of time spent in any state goes to 0.

The contrapositive of the last case shows that there is an invariant distribution, then the markov chain is positive recurrent.

Note that we do not worry about periodicity in CTMCs because there is an implicit self-loop in the "sojourning" time in each state.

Example 5.2 Consider now the random walk reflected at 0 in CTMC land. You go up 1 with rate λ and go

down 1 with rate μ (no self-loops in CTMCs). This gives rise to the following Q matrix:

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ \mu & -(\mu+\lambda) & \lambda & 0 & \dots \\ 0 & \mu & -(\mu+\lambda) & \lambda & \dots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix}$$

This means the jump DTMC will have $P = \Gamma$, i.e. the following:

$$\Gamma = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots \\ \frac{\mu}{\lambda+\mu} & 0 & \frac{\mu}{\lambda+\mu} & 0 & \dots \\ 0 & \frac{\mu}{\lambda+\mu} & 0 & \frac{\mu}{\lambda+\mu} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

It turns out that this is analogous to our random walk reflected at 0 in DTMC land!

- $\lambda < \mu (p < \frac{1}{2})$ (Positive Recurrent)
- $\lambda = \mu (p = \frac{1}{2})$ (Null Recurrent)
- $\lambda > \mu (p > \frac{1}{2})$ (Transient)

Now, let us try to find the invariant distribution in the positive recurrent case. The balance equations read:

$$\begin{aligned} \lambda\pi(i) &= \mu\pi(i+1) \\ \pi(i+1) &= \frac{\lambda}{\mu}\pi(i) \\ &= \rho^{i+1}\pi(0) \end{aligned}$$

Where $\rho = \frac{\lambda}{\mu}$, which is the same as the DTMC equation!

5.5 Lecture 16

We discuss one more discrete-time approximation of a CTMC.

Definition 5.9 (Uniformized DTMC) Fix $\lambda \geq q_i$ for all i, and define

$$P(i,j) = \begin{cases} \frac{q(i,j)}{\lambda} & i \neq j \\ 1 - \frac{q_i}{\lambda} & i = j \end{cases}$$

or equivalently, $P = I + \frac{1}{\lambda}Q$.
Observe that $\pi P = \pi \iff \pi Q = 0$.

We can apply this in the following ways:

- The invariant distribution of X_t can be found by computing P^k as $k \to \infty$.
- Transient distribution of X_t: Define CTMC Y_t with inter-jump times being IID $X \sim \text{Exp}(\lambda)$ and jump probabilities given by P. Then:

$$\pi_t = \sum_{n=0}^{\infty} \pi_0 P^n \frac{(\lambda t)e^{-\lambda t}}{n!}$$

To prove this second claim, we show that our new CTMC Y_t (with self-loops!) has the same behavior as X_t.

Proof Let Z_n be a DTMC with the transition matrix P as defined above.

$$Y_t \equiv_D Z_{N_t}$$

where N_t is a Poisson process with rate λ.

Let T_i be the time spent in state i after entering state i:

$$T_i = J_1 + \cdots + J_S$$

where $J_k \sim \text{Exp}(\lambda)$ (the time until the next jump) and $S \sim \text{Geometric}\left(\frac{q_i}{\lambda}\right)$ (our probability of leaving). From our homework problem (think of Poisson splitting on every geometric), we know that:

$$T_i \sim \text{Exp}(q_i)$$

So, we have proved the sojourning time is the same as the original CTMC. Now we look at the probabilities of leaving. Suppose $i \neq j$

$$\begin{aligned} P(i,j) &= \frac{q(i,j)/\lambda}{\sum_{j \neq i} q(i,j)/\lambda} \\ &= \frac{q(i,j)/\lambda}{q_i/\lambda} = \frac{q(i,j)}{q_i} \end{aligned}$$

Thus, this new CTMC behaves exactly like our old CTMC!

Note that our new CTMC has a litle bit different rate, it is λ.

Here is some intuition about the uniformization of the chain. First, set rates $r_1, \ldots, r_{n-1}$ rates of $j \neq i$ (which should behave like outward flow rates $q(i,j)$). Then, we add a timer for self-loop with rate r_0. Thus, the rate of all jumps is $r_0 + q_i$. This means the number of stays at state i is $\sim$ Geometric $\left(\frac{q_i}{r_0+q_i}\right)$. Furthermore, each stay is $\sim \text{Exp}(r_0 + q_i)$. However, you can prove that the overall stay at i before jumping to $j \neq i$ is $\sim \text{Exp}(q_i)$. Uniformization picks r_i's such that all of this works!

5.5.1 Reversibility

We now talk about reversing Markov chains.

Definition 5.10 (Time-Reversed CTMC) A time-reversed process of some process X_t is:

$$\tilde{X}_t = X_{\tau - t}$$

for some fixed τ.

Theorem 5.11 Assume a CTMC X_t has the invariant distribution π. Then X_t reversed in time is a CTMC with the same invariant distribution, and its matrix $\tilde{Q}$ is given by:

$$\tilde{q}(i,j) = \frac{\pi(j)q(j,i)}{\pi(i)}$$

This suggests the following process:

1. Guess the invariant distribution π for the CTMC under consideration.
2. Guess the CTMC in reversed time and find it's rate matrix $\tilde{Q}$.
3. Show the equation in the theorem above is satisfied.
4. Proves our guesses for the invariant distribution and the cTMC in reversed time are correct.

Definition 5.11 (Reversibility) If a Markov Chain satisfies the detailed balance equations, it is reversible (i.e. $\tilde{q}(i,j) = q(i,j)$). The detailed balance equations for DTMCs is:

$$\pi(i)P(i,j) = \pi(j)P(j,i)$$

for all i, j. The detailed balance equations for a CTMC is:

$$\pi(i)q(i,j) = \pi(j)q(j,i)$$

Intuitively, this means that the chain behaves exactly the same in the reverse direction as it does in the forward direction.

We now present a networks-based application of our CTMCs.

Example 5.3 (M/M/1 Queue) Suppose cusomters arrive at a single-server queue according to a Poisson process with parameter λ, and the server takes time Expo(μ) to handle a packet.
The first M refers to "memoryless" interarrivals, the second refers to "memoryless" services, and 1 represents the amount of simultaneous servers. In general, this can generalize to a A/B/n-type queue which can be first-come first-serve, last-come first-serve, etc. The default is the first one.
We can draw a state transition diagram for the queue as a CTMC; rate λ to go up and μ to go down. Assume $\lambda < \mu$ so the chain is positive recurrent. Then,

- Invariant distribution: $\pi(n) = (1-\rho)\rho^n$ where $\rho = \lambda/\mu$ - we've beaten this horse to death. We can think of ρ as the effective rate of people joining the queue.
- Average number in the queue under the invariant distribution: $\frac{\rho}{1-\rho}$ - can be shown with a straightforward expectation calculation
- Average delay in the system (total time from someone): $\frac{1}{\mu-\lambda}$ - next time

5.6 Lecture 17

Let's prove the claim of average delay $\mathbb{E}[D] = \frac{1}{\mu-\lambda}$ in an $M/M/1$ queue.

Proof First we note that:

$$\mathbb{P}[X_t = k \mid X_{t+\epsilon} = k+1] = \mathbb{P}[X_t = k]$$

this is often termed the Poisson Arrivals See Time Average (PASTA) property. Under invariant conditions, we have that:

$$\mathbb{P}[X_t = k] = \pi(k)$$

So, now we try to find the $\mathbb{E}[D]$:

$$\begin{aligned}
\mathbb{E}[D] &= \sum_{k=0}^{\infty} \frac{k+1}{\mu}\pi(k) \\
&= \sum_{k=0}^{\infty} \frac{k+1}{\mu}(1-\rho)\rho^k \\
&= \frac{1}{\mu}\left(\sum_{k=0}^{\infty} k(1-\rho)\rho^k + \sum_{k=0}^{\infty}(1-\rho)\rho^k\right) \\
&= \frac{1}{\mu}\left(\frac{\rho}{1-\rho} + 1\right) \\
&= \frac{1}{\mu-\lambda}
\end{aligned}$$

from the definition of $\rho = \frac{\lambda}{\mu}$.

Furthermore, it's not hard to see that:

$$D \sim \text{Exp}(\mu - \lambda)$$

Because

$$D = X_1 + \cdots + X_N$$

and $X_i \sim \text{Exp}(\mu)$ and $\mathbb{P}[N = i] = (1-\rho)\rho^i$, $i \geq 0$, where N is independent of X_i (shifted Geometric).

5.6.1 Little's Law and Jackson Network

Now we talk about a brief networking law called Little's Law.

Theorem 5.12 (Little's Law) Suppose we have some black-boxed network with arrival rate μ with average delay through the network D. This means that for L, the average amount of people waiting in the queue, we have $L = \lambda D$.

Let us see an example of this law holding in $M/M/1$ queues.

Example 5.4 We just derived that $L = \frac{\lambda}{\mu-\lambda}$ and $D = \frac{1}{\mu-\lambda}$ so clearly we have $L = \lambda D$ hold.

Intuitively, Little's law is just saying the average occupancy is equal to the average delay times the average arrival rate.

Definition 5.12 (Jackson Networks) Consider some network of J $(\cdot/M/1)$ queues. External arrivals occur according to independent Poisson processes with rate γ_i into queue i. Service time at queue i according to independent exponential distrbituion with rate μ_i. When a customer leaves queue i, independent of the past,

they join queue j with probability $r(i,j)$ and leaves the network with probability $1 - \sum_{j=1}^{J} r(i,j)$. (Each node acts as a $\cdot/M/1$ queue).
Let λ_i be the total rate into a state i. By flow conservation, total arrival rate into queue i is given by

$$\lambda_i = \gamma_i + \sum_{j=1}^{J} \lambda_j r(j,i)$$

for all $i = 1, 2, \ldots, J$.
Thus, we can define $X_t = (X_{1,t}, \ldots, X_{J,t})$ as a multi-dimensional CTMC.

We can find this Markov chain's solution. It's kind of elegant:

Theorem 5.13 (Product Form) Assume that the solution $(\lambda_1, \ldots, \lambda_J)$ is such that $\lambda_i < \mu_i$ (arrival rate < service rate) for all i, then the CTMC X_t admits the following invariant distribution:

$$\pi(x_1, \ldots, x_J) = \pi_1(x_1) \cdot \cdots \cdot \pi_J(x_J)$$

where for each j,

$$\pi_j(n) = (1 - \rho_j)\rho_j^n$$

for $n \geq 0$ and $\rho_j = \frac{\lambda_j}{\mu_j}$, i.e. it acts like J $M/M/1$ queues which are mutually independent!

Example 5.5 Take the following Jackson network where $\lambda < \mu_1, \lambda < \mu_2$.

$$\xrightarrow{\lambda} \mu_1 \longrightarrow \mu_2 \longrightarrow$$

Then:

$$\pi(x_1, x_2) = (1 - \rho_1)\rho_1^{x_1}(1 - \rho_2)\rho_2^{x_2}$$

where $\rho_1 = \frac{\lambda}{\mu_1}$ and $\rho_2 = \frac{\lambda}{\mu_2}$.
What you can do is also guess the reverse-time Markov chain is in the following form:

$$\longleftarrow \mu_1 \longleftarrow \mu_2 \xleftarrow{\lambda}$$

We can find the $\tilde{Q}$ and check if the detailed balance equations are satisfied, or check if the rate matrix condition is satisfied. Then, it shows it's reversible and the invariant distribution is right.

We look at now another Jackson network.

Example 5.6 Suppose now we have the following example queue:

$$\lambda + \gamma p = \gamma \implies \gamma = \lambda/(1 - p)$$

So we must have $\frac{\lambda}{1-p} < \mu$ to have a stable queue. We define $\rho = \frac{\gamma}{\mu} < 1$.

$$\mathbb{P}[X = i] = (1 - \rho)\rho^i$$

5.6.2 CTMC Potpourri Results

We have now a collection about CTMC and Poisson Process facts.

Theorem 5.14 (Residual Time Paradox) Suppose inter-event times are IID non-negative RVs $X_i, i = 1, 2, \ldots$ with PDF $f(x)$ and CDF $F(x)$ and ith moment m_i for $i = 1, 2, \ldots$. Suppose after the process has been running for a long time, an observer arrives at an arbitrary time.

- The inter-event time during which the observer arrives has the PDF $\frac{xf(x)}{m_1}$ (the length of the bus interval). This means its expectation is $\frac{m_2}{m_1}$.
- The residual time to the next event (time I have to wait for the bus) and age time (how long ago the last bus came) from the last event both have the PDF: $\frac{1-F(y)}{m_1}$. The PDF is the same for both because of time-symmetry. Their expectations are $\mathbb{E}[Y] = \frac{m_2}{2m_1}$.

Note that you're more likely to fall in an interval that's bigger rather than smaller; it scales linearly with the PDF by the theorem. Furthermore, the probability that the residual time lies in some interval $(y, y + dy)$ is just $\frac{\mathrm{d}y}{x}$ where x is the length of the bus interval. Multiplying by the PDF and integrating via total probability rule gives our result.

Theorem 5.15 (Sum of independent RVs) Consider random vairables X and Y. Then:

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

i.e. f_Z is the convolution of the two. Furthermore, for characteristic functions,

$$\phi_Z(u) = \phi_X(u)\phi_Y(u)$$

Theorem 5.16 The sum of IID $\sim \mathrm{Exp}(\lambda)$ has the Erlang Distribution.

$$Z \sim \Gamma(n, \lambda);\ f_Z(z) = \frac{\lambda^n z^{n-1} e^{-\lambda z}}{(n-1)!}$$

where $z \geq 0$.

Theorem 5.17 (Poisson Splitting/Thinning) Let $N(t)$ be a Poisson process with rate λ. Each arrival is included in the process $N_1(t)$ if an indpendent $X \sim \mathrm{Bernoulli}(p)$ coin flip results in heads, otherwise it's included in $N_2(t)$. Then, $N_1(t)$ and $N_2(t)$ are independent Poisson processes with rates λp and $\lambda(1-p)$, respectively.

Theorem 5.18 (Poisson Merging) Let $N_1(t), \ldots, N_m(t)$ be m independent Poisson processes with rates $\lambda_1, \ldots, \lambda_m$. Then $N(t) = N_1(t) + \cdots + N_m(t)$ is a Poisson process with rate $\lambda_1 + \cdots + \lambda_m$.

6 Digital Link

6.1 Lecture 18

6.1.1 MAP/MLE

We now turn to the problem of transfering information reliably (i.e. meeting performance requirements) using minimum resources (computation, bandwidth, storage, energy, ...). This is across a physical medium: which could be a phone line cable, fiber or wireless.

To do this, we often have to formulate multiple types of estimation problems. First, we talk about Bayes' Rule. Let $C_1, \dots, C_N$ are "circumstances," while S is a certain "symptom." Observing a particular symptom S, what circumstance happened?

We are given "priors" $p_i = \mathbb{P}[C_i]$ (how frequent in the population that circumstance C_i is) and "conditionals" $q_i = \mathbb{P}[S \mid C_i]$ (how likely you are to have symptom S if you have circumstance C_i).

Theorem 6.1 (Bayes' Rule) Define $\pi_i = \mathbb{P}[C_i \mid S]$. Then $\pi_i = \frac{p_i q_i}{\sum_{j=1}^{N} p_j q_j}$. Note that the bottom probability is just the probability of the symptom, $\mathbb{P}[S]$.

This leads to two ways to estimate what a circumstance you have.

Definition 6.1 (MAP/MLE for events) The Maximum A Posteriori estimate (MAP) is defined as:

$$MAP = \operatorname*{argmax}_i \mathbb{P}[C_i \mid S] = \operatorname*{argmax}_i p_i q_i$$

and the Maximum Likelihood Estimate (MLE) is defined as:

$$MLE = \operatorname*{argmax}_i \mathbb{P}[S \mid C_i] = \operatorname*{argmax}_i q_i$$

Note that if all the priors are the same, then the MAP and MLE are the same. We can also extend this to general discrete random variables:

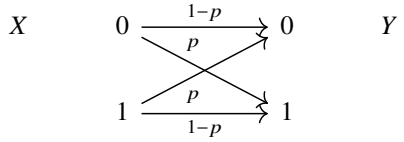
Definition 6.2 (MAP/MLE for Discrete RVs) Let X and Y be discrete random variables.

$$MAP[X \mid Y = y] = \operatorname*{argmax}_x \mathbb{P}[X = x, Y = y]$$

$$MLE[X \mid Y = y] = \operatorname*{argmax}_x \mathbb{P}[Y = y \mid X = x]$$

We think of Y as an output we "observe"; we want to know what input X caused this.

First, we take the case of a Binary Symmetric Channel (BSC).



We transmit either a 0 or 1 from an input X and get a 0 or 1 in the output Y and the channel has a probability p of corrupting the packet. Furthermore, let $\mathbb{P}[X = 1] = \alpha$ (the prior). Then,

Theorem 6.2 For the BSC,

$$MAP[X \mid Y = y] = \begin{cases} \mathbb{1}\{\alpha > (1-p)\} & y = 0 \\ \mathbb{1}\{\alpha > p\} & y = 1 \end{cases}$$

and

$$MLE[X \mid Y] = \begin{cases} Y & p < 0.5 \\ |Y - 1| & p \geq 0.5 \end{cases}$$

where the last expression just flips the observation Y.
We can derive this with a straightforward Bayes Rule calculation using the previous facts. I'll omit the calculations.

Now, we discuss the additive Gaussian Noise channel, $Y = X + Z$ where $Z \sim \mathcal{N}\left(0, \sigma^2\right)$ is independent of X. Suppose $X \in \{0, 1\}$. Let $f_0 = f_{Y|X=0}$ and $f_1 = f_{Y|X=1}$. Furthermore, call the priors p_0, p_1.

Let $\epsilon << 1$, then

$$\begin{aligned} q_0 &= \mathbb{P}\left[Y \in (y, y+\epsilon] \mid X = 0\right] \\ &= f_0(y)\epsilon \\ q_1 &= \mathbb{P}\left[Y \in (y, y+\epsilon] \mid X = 1\right] \\ &= f_1(y)\epsilon \end{aligned}$$

Thus, we can just compare densities.

$$MAP[X \mid Y = y] = \underset{i \in \{0,1\}}{\operatorname{argmax}}\, p_i f_i(y)$$

$$p_1 \frac{e^{\frac{-y-1^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} > p_0 \frac{e^{\frac{-y^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$$

this will simplify algebraically to our result below. The MLE is just comparing q_0 and q_1:

$$MLE[X \mid Y = y] = \underset{i \in \{0,1\}}{\operatorname{argmax}}\, f_i(y)$$

However, by symmetry, the two normal curves intersect at $y = \frac{1}{2}$, so the q_1 over takes q_0 if $y \geq \frac{1}{2}$.

Theorem 6.3 For the Gaussian channel,

$$MAP[X \mid Y = y] = 1\{y \geq \frac{1}{2} + \sigma^2 \ln\left(\frac{p_0}{p_1}\right)\}$$

$$MLE[X \mid Y = y] = 1\{y \geq \frac{1}{2}\}$$

In turn, we can calculate the probability of error under MLE. That is, we transmitted a 0 but got $y \geq \frac{1}{2}$ and misclassified

it as a 1. Call the probability p

$$p = p_0\mathbb{P}[Z \geq 0.5] + p_1\mathbb{P}[Z + 1 < 0.5]$$
$$p = p_0\mathbb{P}[Z \geq 0.5] + p_1\mathbb{P}[Z < -0.5]$$
$$p = p_0\mathbb{P}[Z \geq 0.5] + p_1\mathbb{P}[Z \geq 0.5]$$
$$p = \mathbb{P}[Z \geq 0.5]$$

6.1.2 Huffman Codes

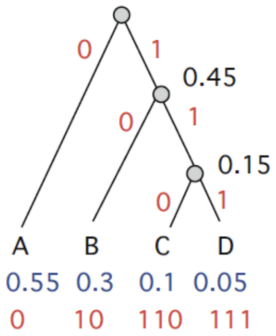
Consider the problem of coding 4 symbols A, B, C, D in a piece of text into binary. One option is assign:

$$A = 00, B = 01, C = 10, D = 11$$

We can decode a received string without errors unambiguously. We have to use two bits per symbol.

Now, suppose we have additional information: the frequency of occurrence of each symbol is 0.55, 0.3, 0.1, 0.05 respectively. Let's assign 0, 10, 110, 111 to A, B, C, D respectively. The reason we are doing this is because we are assigning shorter codes to more frequent symbols. Here, we can also unambiguously decode in one pass since the codes are prefix-free (no code is a prefix of any other). If we compute the average number of bits per symbol, this is now 1.6 (saving us a lot of bits).

This is how we constructed the code. The code is prefix tree because all codewords lie at a leaf node (none is an ancestor of another).



The way we found the codewords is through the following algorithm.

Note 6.1 (Huffman Coding Algorithm) We are given a list of symbols i with probabilities p_i. The Huffman tree is built as follows:

- Take the two symbols with smallest probability a and b.
- Put them as leaves on a tree and combine them into a root $\{a, b\}$ with probability $p_a + p_b$.
- Iteratively continue this, considering $\{a, b\}$ as one character with its new probability (and a, b removed).

Theorem 6.4 The Huffman Code has the smallest average number of bits per symbol among all prefix-free codes.

We will see next lecture why this is the case.

6.2 Lecture 19

6.2.1 Entropy

Let us introduce a new concept of entropy.

Definition 6.3 (Entropy of a Code) Suppose p_i is the probability (frequency) of a symbol i. Then the entropy in a string,

$$H = -\sum_{i=1}^{m} p_i \log_2 p_i$$

You can think of entropy as "average surprise," where we define the surprise from i (or the information revealed by i) as $-\log_2 p_i$. This is because it is a positive quantity which is small when p_i is large and large when p_i is small.
Furthermore, we define $p \log p$ as 0 when p is 0 (this can be seen through L'Hopital's rule).

Let us talk about a specific entropy that is easy to compute.

Example 6.1 We define $H(p) = H$ for a Bernoulli(p). The curve of $H(p)$ vs p looks like an upside-down parabola with maximum at $p = 0.5$ with entropy 1.

Theorem 6.5 For any string of $n \gg 1$ symbols, there are 2^{nH} "typical" strings out of m^n possible strings (where m is the length of the alphabet).

Proof Let X be a random n-length string. Define:

$$\Psi(x) = \frac{1}{n} \log_2 \mathbb{P}[X = x]$$

Note that $\mathbb{P}[X = x] = 2^{n\Psi(x)}$. Let $|x_i|$ denote the number of times symbol i occurs in x and let p_i be the probability of symbol i occurring. Then,

$$\Psi(x) = \frac{1}{n} \log_2 \prod_{i=1}^{m} p_i^{|x_i|}$$

$$\Psi(x) = \sum_{i=1}^{m} \frac{|x_i|}{n} \log_2 p_i$$

However, by the strong law of large numbers, we know $\frac{|x_i|}{n} \to p_i$ a.s. as $n \to \infty$. This means for large n:

$$\Psi(x) = \sum_{i=1}^{m} p_i \log_2 p_i = -H$$

This means that:

$$\mathbb{P}[X = x] = 2^{-nH}$$

for any x, as long as the string follows the strong law of large numbers, i.e. its proportion of characters mirror the true probabilities. Thus, there must be about 2^{nH} typical strings, each with roughly uniform probability.

Example 6.2 Take the binary alphabet, $m = 2$. Then there are $M = 2^n$ possible strings. If $p = 0.1$ (probability of a 1), then $H(p) \approx 0.5$, so the number of typical strings is about $2^{nH} = 2^{0.5n} = \sqrt{M}$.

Now, we will return to source coding.

Theorem 6.6 (Source Coding Theorem) Let γ be the average number of bits per symbol required optimally and let n be the amount of symbols. This means γn is the average number of bits required. This means there are $2^{\gamma n}$ distinct strings, but there are also about 2^{nH} of them. Thus, we have that for any source-coding setup $\gamma = H$.
In other words, H bits per symbol is optimal.

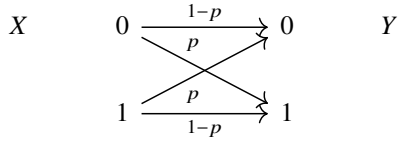
Huffman is not always optimal of all codings (only of the prefix-free ones).

Example 6.3 Consider two symbols A and B where $\mathbb{P}[A] = 0.1$. We again have $H = 0.47$, so this means that the best code uses 0.47 bits/symbol; but Huffman gives 1 bit/symbol (A is 0, B is 1).

6.3 Lecture 20

6.3.1 Information Theory and Channel Coding

We now return to our Binary Symmetric Channel from before.



Definition 6.4 (Rate of Transmission) The rate of information is defined as number of bits conveyed / number of bits transmitted.

For example for our BSC, we could transmit each bit 3 times and have the channel follow majority rule. This leads to a rate of $R = 1/3$. Note that our information-theoretic rate abstracts out the time (this is not megabits per second, for example).

We define a few notions of information theory in order to build one of the most powerful results.

Definition 6.5 Given a channel, let p_e be the maximum probability of an error (over all possible codewords).

Definition 6.6 (Capacity) The capacity of a channel is defined as the fastest rate you can transfer information such that decoding is asymptotically error-free (error rate approaches 0 as $n \to \infty$).

Theorem 6.7 (Capacity of BSC) The capacity of this channel is $C(p) = 1 - H(p)$. where entropy $H(p) = -p \log p - (1-p)\log(1-p)$.

Proof Define E_i to be an indicator for when there is an error in ith position. The sequence of errors/no-errors looks something like $0101100\ldots$ where $\mathbb{P}[E_i = 1] = p$. We can re-use the source-coding theorem from before! For large n the typical amount of these strings is $A = 2^{nH(p)}$.
Suppose there are B different codewords (input strings) of length n that are separable at output. Call them $x_1, \ldots, x_B$ and their outputs the random variables $Y_1, \ldots, Y_B$. Let x_i be composed of n random bits (think of them as vectors). Let S_1 be the set of A typical outputs corresponding to x_1. Then, the probability of an error:

$$\begin{aligned} p_e &= \mathbb{P}[Y_2 \in S_1 \cup Y_3 \in S_1 \cup \cdots \cup Y_B \in S_1] \\ &\leq B \cdot \mathbb{P}[Y_2 \in S_1] \\ &= BA2^{-n} \end{aligned}$$

where the last probability is true because Y_2 has n random bits (since x_2 is arbitrarily random), so $\mathbb{P}[Y_2 \in S_1] = |S_1|2^{-n} = A2^{-n}$. Let $B = 2^{nR}$. Then,

$$\begin{aligned} p_e &\leq 2^{nR}2^{nH(p)}2^{-n} \\ &\leq 2^{n(R+H(p)-1)} \end{aligned}$$

Thus, if we want the probability of an error to go to 0 as $n \to \infty$, then we need $R + H(p) - 1 < 0$ or $R < 1 - H(p)$. We've shown that if $R < C(p)$, then there exists a coding scheme such that $p_e \to 0$ as $n \to \infty$ (it is asymptotically error free). You need the converse to complete the proof: i.e. if $R > C(p)$, then no coding scheme is asymptotically error-free.

Thus we see that our $R = 1/3$ example, there are $B = 2^{n/3}$ different codewords of length n that are separable at output.

In fact, there is a more general theorem for all channels.

Theorem 6.8 (Shannon's Channel Coding Theorem) If the rate of information transfer is below the capacity of a channel, then there exists a channel coding such that error is nominally 0.

Note that neither of these theorems doesn't offer practical coding method for finite n or a required p_e.

6.4 Lecture 21

6.4.1 Hypothesis Testing

We formulate the problem of seeing the output of a link and we want to make predictions of its input. Suppose $X \in \{0, 1\}$ and $\mathbb{P}[Y \mid X]$ is known. Let $\hat{X}$ be the estimate of X. We wish to solve the following optimization problem.

- Maximize Probability of Correct Detection (PCD), $\mathbb{P}\left[\hat{X} = 1 \mid X = 1\right]$
- Subject to a constraint on Probability of False Alarm (PFA), $\mathbb{P}\left[\hat{X} = 1 \mid X = 0\right] \leq \beta$

Note that sometimes in literature, $1-$ PCD (False negative) is called a Type II error and PFA (False positive) is called a Type I error.

Definition 6.7 (Receiver Operating Characteristic (ROC)) This is a graph of PCD for the solution of the above problem as a function $R(\beta)$.

From drawing the curve, we can understand that this function $R(\beta)$ has the following properties:

- $R(\beta)$ is increasing. This is because by being more lenient about PFA, we decide $\hat{X} = 1$ more agressively,
- $R(\beta)$ is concave (i.e. it's second derivative is negative). We get diminishing returns
- $R(1) = 1$, because we can have the decision rule be 1 always
- $R(0) \neq 0$, all $R(0)$ means that you should never show a false positive (but are free to show true positives, i.e. if you knew with 100% certainty what the input is).

Finally, we claim a useful fact that will help us formulate a solution. In particular, we state that the constraint is active at optimum, i.e.

Theorem 6.9 The PFA can't be strictly less than β for the optimal decision rule if the PCD < 1. Suppose for the sake of contradiction that the optimal rule $g_1(y)$ gives:

$$\mathbb{P}\left[\hat{X} = 1 \mid X = 0\right] = \alpha < \beta$$

Then, define another decision rule $g_2(y) = 1$ (it is always 1). Now, consider $g_3(y)$ defined as:

$$g_3(y) = \begin{cases} g_1(y) & \text{with prob. } \frac{1-\beta}{1-\alpha} \\ g_2(y) & \text{otherwise} \end{cases}$$

Now, let us calculate the PFA of this rule:

$$\mathbb{P}[g_3(Y) = 1 \mid X = 0] = \mathbb{P}[g_1(Y) = 1 \mid X = 0]\left(\frac{1-\beta}{1-\alpha}\right) + \mathbb{P}[g_2(Y) = 1 \mid X = 0]\left(1 - \frac{1-\beta}{1-\alpha}\right)$$

$$\mathbb{P}[g_3(Y) = 1 \mid X = 0] = \alpha\frac{1-\beta}{1-\alpha} + 1\left(1 - \frac{1-\beta}{1-\alpha}\right)$$

$$\mathbb{P}[g_3(Y) = 1 \mid X = 0] = \beta \leq \beta$$

Thus, $g_3(y)$ is feasible. Now, let us calculate the objective (PCD). Suppose the PCD of g_1 was $p^* < 1$.

$$\mathbb{P}[g_3(Y) = 1 \mid X = 1] = p^*\left(\frac{1-\beta}{1-\alpha}\right) + 1\left(1 - \frac{1-\beta}{1-\alpha}\right)$$

$$\mathbb{P}[g_3(Y) = 1 \mid X = 1] > p^*$$

which contradicts the optimality of g_1. Thus, we have that the PFA must always be β (except in a degenerate case where we don't have to make prediction).

It turns out that to solve this optimization problem, there is a certain form our decision rule will always have.

Theorem 6.10 (Neyman-Pearson Theorem) To maximize given the constraint, we should predict:

$$\hat{X} = \begin{cases} 1 & \text{if } L(Y) > \lambda \\ 1 & \text{with probability } \gamma, \text{ if } L(Y) = \lambda \\ 0 & \text{if } L(Y) < \lambda \end{cases}$$

where $L(y)$ is the likelihood ratio, i.e. $L(y) = \frac{f_{Y|X}[y|1]}{f_{Y|X}[y|0]}$ and λ, γ are chosen such that $\mathbb{P}\left[\hat{X} = 1 \mid X = 0\right] = \beta$. Note that in the continuous case, the middle probability does not matter.

Let us see Hypothesis testing in action.

Example 6.4 Consider a Gaussian Channel $Y = X + Z$, where $X \in \{0, 1\}$ and $Z \sim \mathcal{N}\left(0, \sigma^2\right)$ is independent of X. The receiver wants to guess X from the received signal Y with $PFA \leq \beta$. In this context,

$$L(y) = \frac{\exp\left(-\frac{(y-1)^2}{2\sigma^2}\right)}{\exp\left(-\frac{y^2}{2\sigma^2}\right)} = \exp\left(\frac{2y-1}{2\sigma^2}\right)$$

where the numerator density is the conditional density, namely the density of $X+Z \sim \mathcal{N}\left(1, \sigma^2\right)$ and denominator is the same thing with mean 0.
Note that $L(y)$ is monotonically increasing and $\mathbb{P}\left[L(y) = \lambda\right] = 0$, so we do not need to worry about the decision rule in this case. Thus, the decision rule will look like:

$$\hat{X} = \begin{cases} 1 & y \geq y_0 \\ 0 & \text{otherwise} \end{cases}$$

for some y_0. Now, we need to choose y_0 such that PFA is β. This is done mathematically.

$$\begin{aligned} \beta &= \mathbb{P}\left[\hat{X} = 1 \mid X = 0\right] \\ &= \mathbb{P}\left[Y > y_0 \mid X = 0\right] \\ &= \mathbb{P}\left[\mathcal{N}(0, \sigma^2) \geq y_0\right] \\ &= \mathbb{P}\left[\mathcal{N}(0, 1) \geq \frac{y_0}{\sigma}\right] \\ &= 1 - \Phi\left(\frac{y_0}{\sigma}\right) \\ y_0 &= \sigma\Phi^{-1}(1-\beta) \end{aligned}$$

Similarly, the PCD can then be calculated knowing y_0.

$$\begin{aligned} \mathbb{P}\left[\hat{X} = 1 \mid X = 1\right] &= \mathbb{P}\left[\mathcal{N}(1, \sigma^2) \geq y_0\right] \\ &= \mathbb{P}\left[\mathcal{N}(0, 1) \geq \frac{y_0 - 1}{\sigma}\right] \\ &= 1 - \Phi(\frac{y_0 - 1}{\sigma}) \end{aligned}$$

By picking different values of β, we can generate a ROC curve. We get different curves for different values of σ^2.

Here is another example.

Example 6.5 (Mean of Exponential RVs.) A machine produces lightbulbs with IID $\text{Exp}(\lambda_x)$ lifespans, where $x \in \{0, 1\}$, $\lambda_0 < \lambda_1$ and $x = 1$ indicates a defective machine. By observing the lifespans of n lightbulbs, we want to detect if the machine is defective with PFA $\leq \beta$.
We observe a vector $Y = (y_1, \ldots, y_n)$. Firstly, we calculate the likelihood ratio,

$$\begin{aligned}
\mathbb{P}[Y] &= \frac{f_{Y|x}(Y \mid 1)}{f_{Y|X}(Y \mid 0)} \\
&= \frac{f_{Y|x}(Y \mid 1)}{f_{Y|X}(Y \mid 0)} \\
&= \frac{\prod_{i=1}^n \lambda_1 \exp(-\lambda_1 y_i)}{\prod_{i=1}^n \lambda_0 \exp(-\lambda_0 y_i)} \\
&= \left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\left(-(\lambda_1 - \lambda_0) \sum_{i=1}^n y_i\right)
\end{aligned}$$

Now note that since $\lambda_1 > \lambda_0$, $L(Y)$ is monotonically decreasing as $\sum_{i=1}^n y_i$ increases. Then, the rule looks something like:

$$\hat{X} = \begin{cases} 1 & \sum_{i=1}^n y_i < a \\ 0 & \text{otherwise} \end{cases}$$

Now we must find a to make the PFA constraint happy. Using the CLT, we sta

$$\begin{aligned}
\beta &= \mathbb{P}\left[\hat{X} = 1 \mid X = 0\right] \\
&= \mathbb{P}\left[\sum_{i=1}^n y_i \leq a \mid X = 0\right]
\end{aligned}$$

This is the CCDF of an Erlang Distribution, but we can make this easier by applying CLT. We see

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^n y_i \leq a\right] &= \mathbb{P}\left[\frac{\sum y_i - \frac{n}{\lambda_0}}{\frac{\sqrt{n}}{\lambda_0}} \leq \frac{a - \frac{n}{\lambda_0}}{\frac{\sqrt{n}}{\lambda_0}}\right] \\
\mathbb{P}\left[\sum_{i=1}^n y_i \leq a\right] &\approx \mathbb{P}\left[\mathcal{N}(0,1) \leq \frac{a\lambda_0 - n}{\sqrt{n}}\right] \\
\beta &= \Phi(\frac{a\lambda_0 - n}{\sqrt{n}}) \\
a &= \frac{n + \sqrt{n}\Phi^{-1}(\beta)}{\lambda_0}
\end{aligned}$$

There is also an example of the discrete case.

Example 6.6 (Discrete Hypothesis Testing) Given $\mathbb{P}[Y \mid X]$ we need a prediction rule for X.

	$Y = A$	$Y = B$	$Y = C$
$X = 0$	0.2	0.5	0.3
$X = 1$	0.2	0.2	0.6

Furthermore, we have the following table for likelihood.

$Y = A$	$Y = B$	$Y = C$
1.0	0.4	2.0

If we arrange the outputs as B, A, C (in increasing $L(y)$), then we have monotonicity. Thus, our decision rule will be a simple threshold. Note that the threshold for $L(y)$, λ, is either $0.4, 1$ or 2, since we can set randomization probability, γ, to be 0, to emulate threshold in between two values.
Suppose we take threshold $\lambda = 1$ with some randomization γ. Then, PCD is:

$$\begin{aligned} PCD &= \mathbb{P}\left[\hat{X} = 1 \mid X = 1\right] \\ &= \mathbb{P}\left[Y = C \mid X = 1\right] + \gamma \mathbb{P}\left[Y = A \mid X = 1\right] \\ &= 0.6 + 0.2\gamma \end{aligned}$$

Now, let us work out the probability of false alarm, PFA.

$$\begin{aligned} PFA &= \mathbb{P}\left[\hat{X} = 1 \mid X = 0\right] \\ &= \mathbb{P}\left[Y = C \mid X = 0\right] + \gamma \mathbb{P}\left[Y = A \mid X = 0\right] \\ &= 0.3 + 0.2\gamma \end{aligned}$$

We can work this out for the other thresholds too. Suppose $\lambda = 2$, then $PCD = 0.6\gamma$ and $PFA = 0.3\gamma$. Suppose $\lambda = 0.4$, then $PCD = 0.8 + 0.2\gamma$ and $PFA = 0.5 + 0.5\gamma$.
Note that together, these PFAs cover the interval $[0, 1]$ with $\gamma \in [0, 1]$. $\lambda = 2$ covers from $[0, 0.3]$, $\lambda = 1$ covers $[0.3, 0.5]$, and $\lambda = 0.4$ covers $[0.5, 1]$. Furthermore, PFA at the boundaries can be met by two different choices; take $\lambda = 1$ and $\gamma = 0$ as well as $\lambda = 2$ and $\gamma = 1$; both give the same PFA and PCD.
This means, given the PFA significance β, choose the corresponding interval for it. Then, we shape the decision rule around that cutoff.

6.5 Lecture 22

6.5.1 Proof of Neyman-Pearson Theorem

We present a proof of the Neyman-Pearson Theorem, now that we've seen it in action. First, we note some intuition. Observe that λ decreases, then we decide $\hat{X} = 1$ more frequently. This means that the PCD and PFA will go up. Then what we do is we set λ, γ such that PFA = β.

Proof 6.1 (Neyman-Pearson) As a reminder, remember our objective.

- Maximize Probability of Correct Detection (PCD), $\mathbb{P}\left[\hat{X} = 1 \mid X = 1\right]$
- Subject to a constraint on Probability of False Alarm (PFA), $\mathbb{P}\left[\hat{X} = 1 \mid X = 0\right] \leq \beta$

Consider some other feasible decision rule $\tilde{X}$ with $\mathbb{P}\left[\tilde{X} = 1 \mid X = 0\right] \leq \beta$ and let the decision rule returned by N-P criterion be $\hat{X}$. We will show that the PCD of $\tilde{X}$ is $\leq$ the PCD of $\hat{X}$.
Now the magical insight. Observe $(\hat{X} - \tilde{X})(L(Y) - \lambda) \geq 0$. To see this, consider three cases:

1. Suppose $L(y) = \lambda$; then the left side is 0 and this is trivially true.
2. Suppose $L(y) > \lambda$; then $\hat{X} = 1 \geq \tilde{X}$ by the N-P decision rule, so both terms in the product are non-negative, meaning the left side is non-negative.
3. Suppose $L(y) < \lambda$; then $\hat{X} = 0 \leq \tilde{X}$ by the N-P decision rule, so both terms in the product are non-positive, meaning the left side is non-negative.

Now, we take the expectation on both sides given the null hypothesis (which we can do since expectation is monotone).

$$\begin{aligned}
\mathbb{E}\left[(\hat{X} - \tilde{X})(L(Y) - \lambda) \mid X = 0\right] &\geq 0 \\
\mathbb{E}\left[\hat{X}L(Y) \mid X = 0\right] - \mathbb{E}\left[\tilde{X}L(Y) \mid X = 0\right] &\geq \lambda(\mathbb{E}\left[\hat{X} \mid X = 0\right] - \mathbb{E}\left[\tilde{X} \mid X = 0\right]) \\
\mathbb{E}\left[\hat{X}L(Y) \mid X = 0\right] - \mathbb{E}\left[\tilde{X}L(Y) \mid X = 0\right] &\geq \lambda(\mathbb{P}\left[\hat{X} = 1 \mid X = 0\right] - \mathbb{P}\left[\tilde{X} = 1 \mid X = 0\right]) \\
\mathbb{E}\left[\hat{X}L(Y) \mid X = 0\right] - \mathbb{E}\left[\tilde{X}L(Y) \mid X = 0\right] &\geq 0 \\
\mathbb{E}\left[\hat{X}L(Y) \mid X = 0\right] &\geq \mathbb{E}\left[\tilde{X}L(Y) \mid X = 0\right]
\end{aligned}$$

The last step is because we assume $\mathbb{P}\left[\hat{X} = 1 \mid X = 0\right] = \beta$ (it is tight) and $\mathbb{P}\left[\tilde{X} = 1 \mid X = 0\right] \leq \beta$.
Furthermore, we note that for any $g(Y)$:

$$\begin{aligned}
\mathbb{E}\left[g(Y)L(Y) \mid X = 0\right] &= \int g(y)L(y)f_{Y|X}(y \mid 0)\,\mathrm{d}y \\
&= \int g(y)\frac{f_{Y|X}(y \mid 1)}{f_{Y|X}(y \mid 0)}f_{Y|X}(y \mid 0)\,\mathrm{d}y \\
&= \int g(y)f_{Y|X}(y \mid 1)\,\mathrm{d}y \\
&= \mathbb{E}\left[g(Y) \mid X = 1\right]
\end{aligned}$$

The same holds for any $g(Y, Z)$ where Z is independent of X, Y. (Think of the Z as the randomization in $\hat{X}$)
Thus, our inequality becomes:

$$\begin{aligned}
\mathbb{E}\left[\hat{X} \mid X = 1\right] &\geq \mathbb{E}\left[\tilde{X} \mid X = 1\right] \\
\mathbb{P}\left[\hat{X} = 1 \mid X = 1\right] &\geq \mathbb{P}\left[\tilde{X} = 1 \mid X = 1\right]
\end{aligned}$$

Thus, $\hat{X}$ is optimal.

6.5.2 Jointly-Gaussian Random Variables

We switch gears to jointly-Gaussian random variables. First, we define a certain notion for multivariate distributions.

Definition 6.8 (Covariance Matrix) The covariance matrix Σ_X of a random variable X is defined as:

$$\Sigma_Y = \mathbb{E}\left[(Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^T\right]$$

Expressing this elementwise is:

$$\begin{aligned}
\Sigma_Y(i,j) &= \text{Cov}\left(Y_i, Y_j\right) \\
\Sigma_Y(i,j) &= \mathbb{E}\left[(Y_i - \mathbb{E}[Y_i])(Y_j - \mathbb{E}\left[Y_j\right])\right] \\
\Sigma_Y(i,j) &= \mathbb{E}\left[Y_i Y_j\right] - \mathbb{E}[Y_i]\,\mathbb{E}\left[Y_j\right] \\
\Sigma_Y(i,i) &= \text{Var}\left(Y_i\right)
\end{aligned}$$

Definition 6.9 (Jointly-Gaussian RVs) $Y_1, \ldots, Y_n$ are jointly-Gaussian random variables if

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = AZ + \mu_Y$$

where Z_i's in $Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_k \end{bmatrix}$ are IID $\mathcal{N}(0,1)$ and A and μ_Y are n-by-k and n-by-1, respectively.

Theorem 6.11 (PDF of Joint-Gaussians) Assuming Σ_Y is nonsingular,

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma_Y}} \exp\left(-\frac{1}{2}(y - \mu_Y)^T \Sigma_Y^{-1}(y - \mu_Y)\right)$$

Theorem 6.12 (Mean and Covariance of J-G) The mean of Y is μ_Y. The covariance of Y is $\Sigma_Y = AA^T$. We write $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$.

Let's prove these two facts.

Proof Recall the change of variables formula. If two random variables U, V are related as $V = aU + b$ with a, b constant, then $f_V(v) = \frac{1}{|a|} f_U(u)$. We can extend this to the vector case. Suppose A, B are constant matrices. Then:

$$V = AU + B \implies f_V(v) = \frac{1}{|\det A|} f_U(u)$$

Consider the PDF of Z:

$$\begin{aligned}
f_Z(z) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right) \\
&= \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{||z||^2}{2}\right)
\end{aligned}$$

Thus, since $Y = AZ + \mu_Y$, we have, $z = A^{-1}(y - \mu_Y)$ (assuming A is invertible). This means that:

$$\begin{aligned}||z||^2 &= ||A^{-1}(y - \mu_Y)||^2 \\ &= (y - \mu_Y)^T \left(A^{-1}\right)^T A^{-1}(y - \mu_Y)\end{aligned}$$

In addition,

$$\begin{aligned}\Sigma_Y &= \mathbb{E}\left[(Y - \mu_Y)(Y - \mu_Y)^T\right] \\ &= \mathbb{E}\left[(AZ)(AZ)^T\right] \\ &= \mathbb{E}\left[AZ^T ZA^T\right] \\ &= A\Sigma_Z A^T \\ &= AA^T\end{aligned}$$

where we realize that $\Sigma_Z = I$ because each Z_i is iid with variance 1. This means that: $\det \Sigma_Y = \det A \cdot \det A^T = (\det A)^2$. This shows that Σ_Y is full rank if and only if A is full rank.

Also,

$$\begin{aligned}(A^{-1})^T A^{-1} &= (A^T)^{-1} A^{-1} \\ &= (AA^T)^{-1} \\ &= \Sigma_Y^{-1}\end{aligned}$$

Thus, by change of variables formula:

$$f_Y(y) = \frac{1}{\det A} \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{||z||^2}{2}\right)$$

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^n |\det \Sigma_Y|}} \exp\left(-\frac{1}{2}(y - \mu_Y)^T \Sigma_Y^{-1}(y - \mu_Y)\right)$$

which is precisely the claim of the PDF.

Note 6.2 Level curves for the Joint-Gaussian PDF are elliptical, with semi-axis lengths σ_i, center at μ_Y, and some axis rotations (because the Σ matrix is not diagonal).

Example 6.7 (2-Dimensional J-Gs) Suppose $\mu_Y = 0$, $k = n = 2$. Then, we can write:

$$\begin{aligned}Y_1 &= a_{11}Z_1 + a_{12}Z_2 \\ Y_2 &= a_{21}Z_1 + a_{22}Z_2\end{aligned}$$

i.e. the independent noises from Z_1, Z_2 is present in both of the Y_i's, so they're not independent. However, it's clear from this expression that the Y_i's are Gaussian (their marginal distribution) because it's a linear combination of two Gaussians.

Example 6.8 Suppose the Y_i's are iid $\mathcal{N}(0, \sigma_i^2)$. Then, we have:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$$

This means the covariance matrix looks like:

$$\Sigma_Y = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

Which means its inverse is:

$$\Sigma_Y^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix}$$

Then, using the form of the joint PDF, we can write:

$$\begin{aligned} f_Y(y_1, y_2) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\begin{bmatrix} y_1 & y_2 \end{bmatrix}\begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right) \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\left(\frac{y_1^2}{\sigma_1^2} + \frac{y_2^2}{\sigma_2^2}\right)\right) \\ &= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{y_1^2}{2\sigma_1^2}\right) \cdot \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{y_2^2}{2\sigma_2^2}\right) \\ &= f_{Y_1}(y_1) f_{Y_2}(y_2) \end{aligned}$$

Theorem 6.13 (Marginal Distribution of J-G RVs) If $Y_1, \ldots, Y_n$ are J-G, then the marginal distribution of $Y_1, \ldots, Y_n$ is Gaussian.

The converse of this does not always hold. You can construct an example where the random variables are marginally Gaussian, but not Jointly Gaussian.

Example 6.9 Consider:

$$f_Y(y_1, y_2) = \frac{1}{\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\left(\frac{y_1^2}{\sigma_1^2} + \frac{y_2^2}{\sigma_2^2}\right)\right)$$

where the density is only defined where $y_1 \cdot y_2 > 0$ (first and third quadrants). Note that this is not Jointly-Gaussian because the level curves cannot be elliptical–namely there is no component in the second and fourth

quadrants. But, we have the following marginal for $y_1 > 0$:

$$\begin{aligned} f_{Y_1}(y_1) &= \frac{1}{\pi\sigma_1\sigma_2}\int_0^\infty \exp\left(-\frac{1}{2}\left(\frac{y_1^2}{\sigma_1^2}+\frac{y_2^2}{\sigma_2^2}\right)\right)\mathrm{d}y_2 \\ &= \frac{1}{\pi\sigma_1\sigma_2}\cdot\frac{1}{2}\sqrt{2\pi}\sigma_2\cdot\exp\left(-\frac{1}{2}\frac{y_1^2}{\sigma_1^2}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{y_1^2}{2\sigma_1^2}\right) \end{aligned}$$

You also have the same marginal for $y_1 \leq 0$, so Y_1 is marginally Gaussian (and so is Y_2, by symmetry). Thus we have found this tricky example.

We also have the following result.

Theorem 6.14 If $Y_1, \ldots, Y_n$ are Jointly-Gaussian RVs, then they are mutually independent if and only if they are pairwise uncorrelated.

Example 6.10 Suppose X and Y and are independent $\mathcal{N}(0,1)$. Consider the random variables $X+Y$ and $X-Y$.

$$\begin{bmatrix} X+Y \\ X-Y \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} X \\ Y \end{bmatrix}$$

This means $X+Y$ and $X-Y$ are jointly-Gaussian.

Theorem 6.15 If V and W are jointly-Gaussian, then their linear combinations $AV+a$ and $BW+b$ are also jointly-Gaussian for some constant matrices A, B and some constant vectors a and b.

7 Tracking

7.1 Lecture 23

7.1.1 LLSE

We state the new problem we attempt to tackle. Let (X, Y) be a pair of continuous RVs related to a system, and we want to estimate X based on the observed value Y by finding a function to estimate it. There are different ways this problem can be presented.

1. The joint distribution of (X, Y) is known.
2. Offline: We observe a set of samples $\{(X_i, Y_i)\}_{i=1}^K$ from the joint distribution.
3. Online: We observe successive samples (X_n, Y_n) and have to iteratively update our guess.

If the prediction is $\hat{X} = g(Y)$, we want to minimize expected cost: $C(g) = \mathbb{E}[c(X, g(Y))]$.

Definition 7.1 (Squared-Error Cost)

$$c(X, \hat{X}) = \left\|X - \hat{X}\right\|^2$$

Definition 7.2 (MMSE) With squared-error cost and minimizing $C(g)$ over an arbitrary function g, $\hat{X}$ is the Minimum Mean Squared Error (MMSE) Estimate of X given Y.

Definition 7.3 (LLSE) With squared-error cost and minimizing $C(g)$ over all affine functions of Y (i.e. $\hat{X} = a + bY$), $\hat{X}$ is the Linear Least Squares Error (LLSE) Estimate of X given Y. We denote this as $L[X \mid Y]$.

It turns out there is a nice closed form for the LLSE.

Theorem 7.1 Assuming $\text{Var}(Y) \neq 0$:

$$L[X \mid Y] = \mathbb{E}[X] + \frac{\text{Cov}(X, Y)}{\text{Var}(Y)}(Y - \mathbb{E}[Y])$$

Proof For LLSE: the cost function is $C(a, b) = \mathbb{E}\left[(X - a - bY)^2\right]$. To optimize this over two variables, we take partial derivatives.

$$\begin{aligned} \frac{\partial}{\partial a} C(a, b) &= 0 \\ 2a - 2\mathbb{E}[X] + 2b\mathbb{E}[Y] &= 0 \\ a &= \mathbb{E}[X] - b\mathbb{E}[Y] \end{aligned}$$

Then,

$$\begin{aligned} \frac{\partial}{\partial b} C(a, b) &= 0 \\ 2b\mathbb{E}\left[Y^2\right] - 2\mathbb{E}[XY] + 2a\mathbb{E}[Y] &= 0 \\ b &= \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} \end{aligned}$$

Combining things proves the claim.

Example 7.1 Suppose we knew $Y = \alpha X + Z$ where X,Z are zero-mean and independent.

$$\begin{aligned}\text{Cov}(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[(X(\alpha X + Z))] \\ &= \alpha\mathbb{E}\left[X^2\right] + \mathbb{E}[XZ] \\ &= \alpha\mathbb{E}\left[X^2\right]\end{aligned}$$

Furthermore,

$$\begin{aligned}\text{Var}(Y) &= \alpha^2\text{Var}(X) + \text{Var}(Z) \\ &= \alpha^2\mathbb{E}\left[X^2\right] + \mathbb{E}\left[Z^2\right]\end{aligned}$$

This means that the LLSE is:

$$\begin{aligned}L[X \mid Y] &= \frac{\alpha\mathbb{E}\left[X^2\right]}{\alpha^2\mathbb{E}\left[X^2\right] + \mathbb{E}\left[Z^2\right]}Y \\ &= \frac{\alpha^{-1}Y}{1 + SNR^{-1}}\end{aligned}$$

where SNR is the signal-to-noise ratio, i.e.:

$$SNR = \frac{\alpha^2\mathbb{E}\left[X^2\right]}{Z^2}$$

Note that if SNR is high, then $\mathbb{L}[[]\, X \mid Y] = \alpha^{-1}Y$ and if SNR is low, then $L[X \mid Y] = \mathbb{E}[X] = 0$ (observing Y doesn't help use find X).

We define the signal-to-noise ratio more generally.

Definition 7.4 (Signal-to-Noise Ratio) Suppose you have some signal W and noise Z. $\mathbb{E}\left[W^2\right]$ represents signal power and $\mathbb{E}\left[Z^2\right]$ represents noise power. This means that the signal to noise ratio is:

$$SNR = \frac{\mathbb{E}\left[W^2\right]}{\mathbb{E}\left[Z^2\right]}$$

7.2 Lecture 24

7.2.1 Geometry of LLSE

We can always think of the LLSE $L[X \mid Y]$ as the projection of X onto the subspace $\mathcal{L}(Y)$ of linear functions of Y. To do this, first we have to set up the idea of a Hilbert space.

Definition 7.5 (Vector Space) A set S is a vector space if it's closed under addition and scalar multiplication (among other properties).

Definition 7.6 (Hilbert Space) A set $\mathcal{H}$ is a Hilbert space if it's a vector space that has an inner product and is complete (all limits of convergent sequences are in the space).

Definition 7.7 (Inner Product) An (real) inner product on a vector space S is an operation $\langle \cdot, \cdot \rangle : S \to \mathbb{R}$ that satisfies.

Definition 7.8 (Norm) The norm or length of a vector v given an inner product is $\|v\| = \sqrt{\langle v, v \rangle}$.

We now extend our optimization-motivated LLSE in order to allow our estimate to be a general function, i.e. the MMSE. First, we claim that the LLSE $L[X \mid Y]$ is the "projection" of X onto the subspace of affine functions of Y, i.e. $\mathcal{L}(Y) = \{c + dY \mid c, d \in \mathbb{R}\}$. This is a vector space, but we need an inner product to make it a Hilbert space. For any space of RVs, let us define the inner product as:

$$\langle V, W \rangle = \mathbb{E}[VW]$$

We define orthogonality based on if the inner product is 0, i.e. two random variables V and W are orthogonal if $\mathbb{E}[VW] = 0$.

Theorem 7.2 (Hilbert's Projection Theorem) Consider some Hilbert space $\mathcal{H}$ and $a \in \mathcal{H}$ and some subspace U of $\mathcal{H}$. Then, there exists a linear projection operator P that projects onto U (finds the element of U, u^*, such that $||a - u^*||$ is minimized), so the projection of is Pa. Then, we claim that the error, $a - Pa \in U^\perp$, i.e. for any $u \in U$, $\langle a - Pa, u \rangle = 0$.

In particular, our claim of LLSE projection can be rewritten as:

Theorem 7.3 (Projection Property) $X - L[X \mid Y] \in \mathcal{L}(Y)^\perp$, where $Z^\perp$ denotes the orthogonal complement to vector space Z. In particular, this means it is orthogonal to every affine function of Y.

Proof Call $L[X \mid Y] = a + bY$. Consider $c + dY \in \mathcal{L}(Y)$. Then, we will show $\mathbb{E}[(X - a - bY)(c + dY)] = 0$. We know from our partial derivatives before, in the proof of Theorem 7.1,

$$\mathbb{E}[X - a - bY] = 0$$
$$\mathbb{E}[Y(X - a - bY)] = 0$$

Thus, taking c times the the first equation and adding d times the second equation shows the claim.

This implies, from the Hilbert Projection Theorem, that $L[X \mid Y]$ is the closest point to X in $\mathcal{L}(Y)$. However, we can prove this without invoking the result.

Proof We will show $\mathbb{E}\left[(X - L[X \mid Y])^2\right] \leq \mathbb{E}\left[(X - h(Y))^2\right]$ for any $h(Y) = c + dY$. To do this, note that:

$$\begin{aligned}\mathbb{E}\left[(X - h(Y))^2\right] &= \mathbb{E}\left[(X - L[X \mid Y] + L[X \mid Y] - h(Y))^2\right] \\ &= \mathbb{E}\left[(X - L[X \mid Y])^2\right] + \mathbb{E}\left[(L[X \mid Y] - h(Y))^2\right] + 2\mathbb{E}\left[(X - L[X \mid Y])(L[X \mid Y] - h(Y))\right]\end{aligned}$$

Note that $L[X \mid Y] - h(Y) \in \mathcal{L}(Y)$, so it must be orthogonal to the error. This means the cross term is 0. This yields:

$$\mathbb{E}\left[(X - h(Y))^2\right] \geq \mathbb{E}\left[(X - L[X \mid Y])^2\right]$$

with equality if the other square term is 0, i.e. $h(Y) = L[X \mid Y]$.

Now remember our online/offline regression problem; we can replace all the statistics by their sample estimates (like sample average). By law of large numbers, as the number of samples increases, the linear regression approaches the LLSE estimate.

7.3 MMSE

We now consider the MMSE problem. We want to find the function g such that $g(Y)$ minimizes $\mathbb{E}\left[(X - g(Y))^2\right]$.

Theorem 7.4 (MMSE) MMSE of X given Y is given by $g(Y) = \mathbb{E}[X \mid Y]$.

Remember that $\mathbb{E}[X \mid Y]$ is a random variable, and $\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x$ where $f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$.

Theorem 7.5 (Orthogonality of MMSE)

- For any function $\phi(\cdot)$, $\mathbb{E}[(X - \mathbb{E}[X \mid Y])\phi(Y)] = 0$.
- If $g(Y)$ is such that $X - g(Y)$ is orthogonal to any function of Y, $g(Y) = \mathbb{E}[X \mid Y]$ (i.e. it's the unique function with this property).

Proof We start with the first claim. Consider the following tower rule expression with arbitrary $\phi(Y)$:

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X \mid Y]\,\phi(Y)] &= \int_{-\infty}^{\infty} \mathbb{E}[X \mid Y = y]\,\phi(y) f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \frac{f_{X,Y}(x,y)}{f_Y(y)} \phi(y) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\phi f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \mathbb{E}[x\phi(y)]\end{aligned}$$

Thus, this means that $\mathbb{E}[X\phi(Y) - \mathbb{E}[X \mid Y]\,\phi(Y)] = 0$, as we wanted to show.
The second claim is that of uniqueness. Consider some $g(Y)$ which has $X - g(Y)$ orthogonal to any function of Y. We will show the norm of $g(Y) - \mathbb{E}[X \mid Y]$ is 0, meaning they must be the same random variable almost surely.

$$\begin{aligned}\mathbb{E}\left[(g(Y) - \mathbb{E}[X \mid Y])^2\right] &= \mathbb{E}[(g(Y) - \mathbb{E}[X \mid Y])((g(Y) - X) - (\mathbb{E}[X \mid Y] - X))] \\ &= \mathbb{E}[(g(Y) - \mathbb{E}[X \mid Y])((g(Y) - X)] - \mathbb{E}[(g(Y) - \mathbb{E}[X \mid Y])(\mathbb{E}[X \mid Y] - X)]\end{aligned}$$

Call the first term T_1, the second term T_2. $\mathbb{E}[T_1] = 0$ because its left term is a function of Y and $g(Y) - X$ is error, so they must be orthogonal. Furthermore, $\mathbb{E}[T_2] = 0$ because its left term is a function of Y and $\mathbb{E}[X \mid Y] - X$ is error, so they must be orthogonal.

By the Hilbert projection theorem, this theorem shows the form of the MMSE as $\mathbb{E}[X \mid Y]$. You can prove this similarly to the LLSE case using the Pythagorean Theorem.

Here is a lot of properties of conditional expectation.

Theorem 7.6 (Properties of MMSE)

- Linearity: $\mathbb{E}[a_1X_1 + a_2X_2 \mid Y] = a_1\mathbb{E}[X_1 \mid Y] + a_2\mathbb{E}[X_2 \mid Y]$
- Factoring: $\mathbb{E}[h(Y)X \mid Y] = h(Y)\mathbb{E}[X \mid Y]$
- Independence: If X and Y are independent, $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$
- Smoothing: $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$
- Tower: $\mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Y] = \mathbb{E}[X \mid Y]$

Let's prove the last of them, which the most tricky of the bunch.

Proof Let $\mathbb{E}[X \mid Y, Z] = R$, where $R \in \mathcal{G}(Y, Z)$. We want to show the error of the estimation is 0.

$$\mathbb{E}[(R - \mathbb{E}[X \mid Y])\phi(Y)] = \mathbb{E}[(X - \mathbb{E}[X \mid Y])\phi(Y)] - \mathbb{E}[(X - R)\phi(Y)]$$

Clearly the first term is 0 from our orthogonality lemma. Note that $X - R$ is the error from $\mathbb{E}[X \mid Y, Z]$, but $\phi(Y) \in \mathcal{G}(Y) \subseteq \mathcal{G}(Y, Z)$, so the second term is also 0. This means that the error is indeed 0, meaning that $R - \mathbb{E}[X \mid Y]$ is orthogonal to Y. meaning that $\mathbb{E}[R \mid Y] = \mathbb{E}[X \mid Y]$.

Let us do some examples to practice the properties.

Example 7.2 Let X, Y be IID $U[0, 1]$. We wish to find $\mathbb{E}\left[(X + 2Y)^2 \mid Y\right]$.

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

Example 7.3 Let X, Y, Z be IID. We want to compute $e = \mathbb{E}[X \mid X + Y + Z]$. Notice by symmetry,

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

This means that $\mathbb{E}[(X + Y + Z) \mid (X + Y + Z)] = 3e$. Thus, $3e = (X + Y + Z)$. Our answer is $e = \frac{X+Y+Z}{3}$.

Returning to our Jointly-Gaussian RVs, we have an even cleaner result.

Theorem 7.7 Let X, Y be JG RVs. Then,

$$\mathbb{E}[X \mid Y] = LX \mid Y = \mathbb{E}[X] + \frac{\text{Cov}(X, Y)}{\text{Var}(Y)}(Y - \mathbb{E}[Y])$$

Proof We first note that $X - L[X \mid Y]$ and Y are uncorrelated, because:

$$\begin{aligned}\text{Cov}(() X - L[X \mid Y], Y) &= \mathbb{E}[(X - L[X \mid Y])Y] - \mathbb{E}[X - L[X \mid Y]]\,\mathbb{E}[Y] \\ &= 0 - (\mathbb{E}[X] - \mathbb{E}\left[\mathbb{E}[X] + \frac{\text{Cov}(X, Y)}{\text{Var}(Y)}(Y - \mathbb{E}[Y])\right])\mathbb{E}[Y] \\ &= 0 \cdot \mathbb{E}[Y] = 0\end{aligned}$$

Also, then $X - L[X \mid Y]$ and Y are JG, as they are linear combinations of X and Y. This means that $X - L[X \mid Y]$

and Y are independent. This means that:

$$\begin{aligned}\mathbb{E}\left[X - L[X \mid Y] \mid Y\right] &= \mathbb{E}\left[X - L[X \mid Y]\right] \\ &= 0\mathbb{E}\left[X \mid Y\right] \\ &= L[X \mid Y]\end{aligned} \qquad = \mathbb{E}\left[L[X \mid Y] \mid Y\right]$$