EECS 126

Probability and Random Processes

Lecture Notes

Druv Pai

Contents

1 Introduction

The professor is Shyam Parekh, whose office hours are Tuesday, from 2 PM to 3 PM, in 258 Cory Hall.

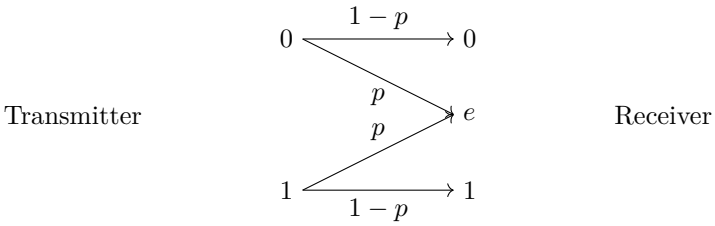
In the beginning, probability was very ad-hoc and application driven, but later it was rigorized and backed by formalization. So we'll try to cover a mixture of both.

We'll start with some theory and mechanics of probability spaces and random variables on these probability spaces, e.g. Bayes' Law, expectations, and distributions, and information theory.

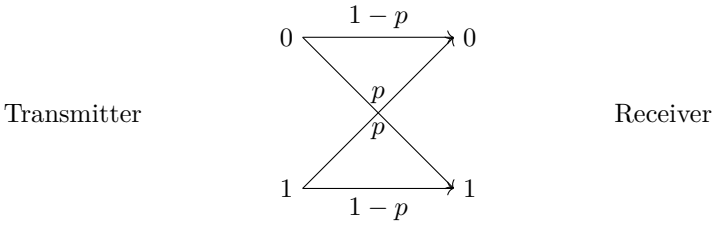
Then, we'll move into applications, including Markov chains, binary erasure channel, and Kalman/EM filter.

Example Application

One application we'll talk about is the capacity of the binary erasure channels, which is related to information theory and communication.



The rate r of this binary erasure channel is given by, where the number of bits of information L in a string that is n bits long, $r = L/n$. The capacity is the largest achievable rate such that $\lim_{n\to\infty} \mathbb{P}[\text{error}] = 0$. Another example of a binary erasure channel is a channel where bits can either be transmitted successfully or flipped: a binary symmetric channel.



Another application we consider is an Erdös-Renyi random graph. Say we have a graph on n vertices, and $p(n)$ be a probability of an edge present between a given pair (where the probability that a given edge exists is independent

of any other edge). If $p(n) = \lambda \log(n)/n$ and $\lambda > 1$, $G(n, p(n))$ is connected almost surely (with probability 1) as n grows unboundedly. On the other hand, if $\lambda < 1$, then $G(n, p(n))$ is disconnected almost surely as n grows unboundedly. This is a sharp boundary and there's very little room for error on λ; if $\lambda = 1$, then the probability of connectedness is a complicated function of n.

Another application we consider is the Markov chain. In a Markov chain, the current state is sufficient information to evaluate probabilities at all points in the future. Markov chains can have finite or infinite state space (both countable and uncountable), and be discrete time or finite time. One instance where we use Markov chains are queueing theory. We also can introduce the notion of reversibility of a Markov chain, to extract past information from the present state.

Another application is estimation given data, specifically the case of maximum likelihood estimation (MLE); we want to choose the value that maximizes the probability of observed data:

$$\hat{\theta}_{\text{MLE}} = \operatorname*{argmax}_{\theta} \mathbb{P}[\mathcal{D} \mid \theta]$$

Another estimator is the maximum a posteriori estimator (MAP). We want to choose the value that is "most" probable given an observed data set $\mathcal{D}$ and a prior belief as to the distribution of the estimator parameters:

$$\hat{\theta}_{\text{MAP}} = \operatorname*{argmax}_{\theta} \mathbb{P}[\theta \mid \mathcal{D}] = \operatorname*{argmax}_{\theta} \mathbb{P}[\mathcal{D} \mid \theta]\mathbb{P}[\theta] = \frac{\mathbb{P}[\theta \cap \mathcal{D}]}{\mathbb{P}[\mathcal{D}]}$$

In this case the $\mathbb{P}[\theta]$ is the prior concerning the estimator parameters.

Another relevant application is hypothesis testing. If $X = 0$ or 1, then this indicates whether there actually is no fire, or there is a fire. If $\hat{X} = 0$ or 1, this indicates whether we have determined whether there is a fire or not. We want to minimize the probability of misdetection $\mathbb{P}\left[\hat{X} = 0 \mid X = 1\right]$ subject to the probability of false alarm being constrained: $\mathbb{P}\left[\hat{X} = 1 \mid X = 0\right] \leq \beta$. This gives rise to the *Neyman-Pearson Theorem.*

The Kalman Filter is a robust solution to the problem of online estimation. We are given a classical control diagram:

$$\begin{aligned} x(n+1) &= Ax(n) + v(n), \quad n \geq 0 \\ y(n) &= Cx(n) + w(n) \end{aligned}$$

where v and w are noise functions. We solve this using probabilistic control theory models.

2 Probabilistic Models

The first thing we want to do when talking about probabilistic models is to

- define an experiment
- determine the sample space Ω of all possible outcomes
- find the probability law that assigns probability to subsets of Ω outcomes

Example 1. We flip a coin once. The experiment is flipping a coin, $\Omega = \{H, T\}$, and the probability law is: $\mathbb{P}[\Omega] = 1$, $\mathbb{P}[H] = p$, $\mathbb{P}[T] = 1 - p$, and $\mathbb{P}[\emptyset] = 0$. This looks easy but can get very hard.

It may be best to consider Ω as a set of disjoint outcomes, and a set of outcomes is an event. The probability law maps events to probabilities.

Definition 2 (σ-Algebra). Under measure theory, we define a σ-algebra $\mathbb{F}$ on a set X as a set of subsets of X that has the following properties:

- $X \in \mathbb{F}$ (the "universal" set)
- $S \in \mathbb{F} \implies S^c \in \mathbb{F}$
- $X_1, \ldots, \in \mathbb{F} \implies \bigcup_i X_i \in \mathbb{F}$

Some σ-algebras can also be equipped with a measure μ.

Remark 3. In our regular probability theory, we have that our σ-algebra is the power set of Ω i.e. $\mathbb{F} = \text{pow}(\Omega)$ and our probability measure is the probability function Pr.

Example 4. Take $[0,1]$; for any $x \in [0,1]$, $\mu(x) = 0$. Indeed, $\mu(\mathbb{Q}[0,1]) = 0$. In general, any set of countable cardinality in an uncountable cardinality sample space has measure 0; some uncountable cardinality sets in an uncountable cardinality sample space have measure 0 as well, like the Cantor set.

Lemma 5. The probability law Pr (probability measure on the σ-algebra $\text{pow}(\Omega)$) must satisfy

- non-negativity: $\mathbb{P}[A] \geq 0$ for each event A
- additivity: for a sequence of disjoint events A_i, we have $\mathbb{P}[\bigcup_i A_i] = \sum_i \mathbb{P}[A_i]$
- normalization: $\mathbb{P}[\Omega] = 1$

due to the earlier definitions.

It's necessary at this point to cover the algebra of sets; first, is distributivity:

$$S \cup (T \cap U) = (S \cup T) \cap (S \cup U)$$

and

$$S \cap (T \cup U) = (S \cap T) \cup (S \cap U)$$

We also have DeMorgan's laws:

Theorem 6 (DeMorgan). Complements distribute:

$$\left(\bigcup_i S_i\right)^c = \bigcap_i S_i^c$$

and

$$\left(\bigcap_i S_i\right)^c = \bigcup_i S_i^c$$

From this we get further properties of the probability law:

- If $A \subseteq B$, then $\mathbb{P}[A] \leq \mathbb{P}[B]$.
- Principle of Inclusion-Exclusion: $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$.
- $\mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$.
- Another form of the PIE: $\mathbb{P}[A \cup B \cup C] = \mathbb{P}[A] + \mathbb{P}[A^c \cap B] + \mathbb{P}[A^c \cap B^c \cap C]$.

Finally we look at conditional probability:

Example 7. Consider a fair dice (that is, where $\mathbb{P}[i] = 1/6$ for $i = 1, \ldots, 6$). We have a sample space with 6 outcomes; we associate a probability measure Pr with each outcome, which happens to be 1/6. We want to find

$$\mathbb{P}[\text{outcome is 6} \mid \text{outcome is even}]$$

We partition the sample space into two categories: the outcome being odd or even. Since there are three even outcomes, and there is one outcome with value 6, and the outcomes are equiprobable, the desired probability is 1/3.

The way we formalize this is through Bayes' Law.

Theorem 8 (Bayes' Law). For events A and B, if $\mathbb{P}[B] \neq 0$, then

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

Let's apply this to our example. If A is the event "outcome is 6" and B is the event "outcome is even", then $\mathbb{P}[A \cap B] = \mathbb{P}[A] = 1/6$, and $\mathbb{P}[B] = 1/2$, and $(1/6)/(1/2) = 1/3$, confirming our claim.

The conditional probability measure also follows the rules of non-negativity, additivity, and normalization.

Example 9. Say we have two rolls of a fair four-sided die (that is, $\mathbb{P}[i] = 1/4$ for $i = 1, \dots, 4$). The sample space looks like a 4×4 grid:

	1	2	3	4
1	1/16	1/16	1/16	1/16
2	1/16	1/16	1/16	1/16
3	1/16	1/16	1/16	1/16
4	1/16	1/16	1/16	1/16

Define event B as $\min(X, Y) = 2$, if X is the random variable that denotes the value of the first roll and Y is the random variable that denotes the value of the second roll. Outcomes that are in B are colored blue.

Define A_m as $\max(X, Y) = m$. Then from the grid, there are 5 blue points, then there is 1 point where $X = Y = 2$, so $\mathbb{P}[A_2 \mid B] = 1/5$; there are 2 blue points where $X = 3$, and 2 if $X = 4$, but none if $X = 1$, so $\mathbb{P}[A_3 \mid B] = \mathbb{P}[A_4 \mid B] = 2/5$ and $\mathbb{P}[A_1 \mid B] = 0$.

Example 10. If an airplane is present, it's detected with probability 0.99. Define the probability of raising an alarm when no airplane is present, as the probability of false alarm, and let this be 0.1. The probability of a plane being present at any given time is 0.05.

We want to compute:

1. the probability of no airplane present and a false alarm

   Solution. Define A as the event that an airplane is present, and B is the probability that the alarm is raised. We want to compute $\mathbb{P}[A^c \cap B]$. We know that $\mathbb{P}[B \mid A^c] = 0.1$, and $\mathbb{P}[B \mid A] = 0.99$, and $\mathbb{P}[A] = 0.05$. Then we have
   $$\mathbb{P}[A^c \cap B] = \mathbb{P}[B \mid A^c]\mathbb{P}[A^c] = 0.1 \cdot (1 - 0.05) = 0.095$$
   as desired. □

2. The probability of an airplane present and no detection

   Solution. We use the same events as before. Then we're looking for
   $$\mathbb{P}[A \cap B^c] = \mathbb{P}[B^c \mid A]\mathbb{P}[A] = (1 - \mathbb{P}[B \mid A])\mathbb{P}[A] = (1 - 0.99) \cdot 0.05 = 0.0005$$
   as desired. □

Example 11. Say we have three cards drawn randomly from a deck of 52 cards *without replacement* (that is, after drawing a card we don't place them back into the deck). We want to find the probability that no heart is drawn in the three cards.

Solution. Define S_i as the event that the ith card is not a heart. Then we're computing, by Principle of Inclusion-Exclusion,
$$\mathbb{P}[S_1 \cap S_2 \cap S_3] = \mathbb{P}[S_1]\mathbb{P}[S_2 \mid S_1]\mathbb{P}[S_3 \mid (S_1 \cap S_2)] = \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{37}{50}$$
as desired. □

Theorem 12 (Total Probability Theorem). Let $A_1, \dots, A_n \subseteq \Omega$ be disjoint events that form a *partition* of Ω (that is, $\bigcup_{i=1}^n A_i = \Omega$). Let $B \subseteq \Omega$ be an event. Then
$$\mathbb{P}[B] = \sum_{i=1}^n \mathbb{P}[B \cap A_i] = \sum_{i=1}^n \mathbb{P}[B \mid A_i]\mathbb{P}[A_i]$$

A theorem that follows from that is a modification of Bayes' Rule:

Theorem 13 (Bayes' Rule Modification). Let $A_1, \ldots, A_n \subseteq \Omega$ be disjoint events that form a partition of Ω. Assume $\mathbb{P}[A_i] > 0$ for each i. Then for any event $B \subseteq \Omega$ such that $\mathbb{P}[B] > 0$, we have

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

and, by the Total Probability Theorem, we have

$$\mathbb{P}[A_i \mid B] = \frac{\mathbb{P}[B \mid A_i]\mathbb{P}[A_i]}{\sum_{j=1}^{n} \mathbb{P}[B \mid A_j]\mathbb{P}[A_j]}$$

This theorem relates, probabilistically, causes to effects. A_i are a set of causes, and B is an effect; we can find the **posterior probability** $\mathbb{P}[A_i \mid B]$ given the **prior probability** $\mathbb{P}[A_i]$ (which is an assumption on the event A_i), and the **new knowledge** $\mathbb{P}[B \mid A_i]$. In essence, we update our knowledge of the causes based on the effects.

Example 14 (False Positive Puzzle). Consider a test for a rare disease. A random person has the disease with probability 0.001.

- If a person has the disease, the test is positive with probability 0.95.
- If a person doesn't have the disease, the test is negative with probability 0.95.

Given that the test is positive, what is the probability that the person has the disease?

Solution. Let A be the event that the person has the disease, and B is the event that the test comes back positive. We want to find $\mathbb{P}[A \mid B]$. By Bayes' Rule, we have

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[B \mid A]\mathbb{P}[A]}{\mathbb{P}[B \mid A]\mathbb{P}[A] + \mathbb{P}[B \mid A^c]\mathbb{P}[A^c]} = \frac{0.95 \cdot 0.001}{0.95 \cdot 0.001 + 0.05 \cdot 0.999} \approx 0.018664$$

which is less than 2%, a lot less than what was expected. □

What's really happening here? Let's try to build some intuition. One way to see this result intuitively is to see that most of the sample space is taken by A^c, the event where the person does not have the disease. When we're in A, most of the area is taken by B. And within A^c, most of the area is taken by B^c. But because there are so many points in A^c, even the space that is not taken by B^c is larger than the space of B within A. Basically, 5% of a large number is a lot larger than 95% of a small number. Formally, $\mathbb{P}[B \cap A^c] = 0.04995 \gg 0.00095 = \mathbb{P}[B \cap A]$.

In general, if $x \gg y$, then $x/y \gg 1$, then $x/y + 1 \gg 1$, then $(x+y)/y \gg 1$, then $y/(x+y) \ll 1$.

Definition 15 (Independence of Events). Event A is independent of event B if and only if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$.

Lemma 16. The following are implications of independence:

- If A is independent of B, then B is independent of A, and A and B are independent.
- If $\mathbb{P}[B] > 0$, then $\mathbb{P}[A \mid B] = \mathbb{P}[A]$ if and only if A and B are independent.
- If A and B are disjoint, then $\mathbb{P}[A \cap B] = 0$, thus $\mathbb{P}[A \cap B] \neq \mathbb{P}[A]\mathbb{P}[B]$, so A and B cannot be independent unless $\mathbb{P}[A] = 0$ or $\mathbb{P}[B] = 0$.

Example 17. Consider two rolls of a fair 4-sided die. Each of 16 outcomes is equally likely. Define A_i as the event that the first roll is i and B_j as the event that the second roll is j. Define C as the event that the first roll is 1 and D as the event that the sum of the two rolls are 5. Define E as the event that the maximum of the two rolls is 2 and F as the event that the maximum of the two rolls is 2.

Claim 18.

1. Independence of A_i and B_j:

Solution. There are 4 equiprobable outcomes for the first row and so $\mathbb{P}[A_i] = 1/4$, and likewise for B_j. Also $A_i \cap B_j$ is a single outcome in the sample space, and since they are equiprobable we have that $\mathbb{P}[A_i \cap B_j] = 1/16 = (1/4)^2 = \mathbb{P}[A_i]\mathbb{P}[B_j]$, thus A_i and B_j are independent. □

2. Independence of C and D:

   Solution. Clearly by the first part we know that $\mathbb{P}[C] = 1/4$ and by listing out the possibilities for the second roll, we know that there are four of them, and the intersection of the two rolls is an equiprobable point in the sample space, so $\mathbb{P}[D] = 1/4$. Clearly $\mathbb{P}[C \cap D] = 1/16 = \mathbb{P}[C]\mathbb{P}[D]$. □

3. Non-independence of E and F:

   Solution. Clearly $\mathbb{P}[E \cap F] = 1/16$ since both rolls need to be 2. Meanwhile, by counting on the sample space, there are 3 events in which the maximum of the two rolls is 2, so $\mathbb{P}[E] = 3/16$. Similarly, $\mathbb{P}[F] = 5/16$, and so the products do not equal, hence independence does not hold. □

Definition 19 (Conditional Independence). Given an event C, the events A and B are **conditionally independent** if and only if $\mathbb{P}[A \cap B \mid C] = \mathbb{P}[A \mid C]\mathbb{P}[B \mid C]$.

Note that unconditional independence of the events A and B does not imply conditional independence with respect to an event C, and vice versa.

Definition 20 (Independence of Several Events). The events $A_1, \dots, A_n$ are independent if

$$\mathbb{P}\left[\bigcap_{i \in S} A_i\right] = \prod_{i \in S} \mathbb{P}[A_i]$$

for every set $S \subseteq \{1, \dots, n\}$.

Definition 21 (Mutual Independence). The events $A_1, \dots, A_n$ are **independent** if

$$\mathbb{P}\left[\bigcap_{i=1}^{n} A_i\right] = \prod_{i=1}^{n} \mathbb{P}[A_i]$$

Example 22 (Pairwise Independence Doesn't Imply Mutual Independence). Consider two fair coin tosses. Let H_1 be the event that the first toss is a head, H_2 be the event that the second toss is a head, and D be the event that the two tosses have different outcomes. Obviously H_1 and H_2 are pairwise independent.

Claim 23.

1. D and H_1 are pairwise independent, and D and H_2 are pairwise independent.

   Solution. We compute $\mathbb{P}[D \mid H_1] = \mathbb{P}[D \cap H_1]/\mathbb{P}[H_1] = (1/4)/(1/2) = 1/2 = \mathbb{P}[D]$, so D and H_1 are independent. Similarly D and H_2 are independent. □

2. D and H_1 and H_2 are not mutually independent.

   Solution. Consider $\mathbb{P}[D \cap H_1 \cap H_2] = 0$ since we cannot have a different outcome and both H_1 and H_2 occur. But none of the probabilities of the events are 0, so we do not have mutual independence. □

Example 24 (Mutual Independence Doesn't Imply Pairwise Independence). Consider two independent rolls of a fair 6 sided die. Let A be the event that the first roll is 1, 2, or 3, let B be the event that the first roll is 3, 4, or 5, and let C be the event that the sum of the two rolls is 9.

Claim 25.

1. A, B, and C are mutually independent.

   Solution. $\mathbb{P}[A]\mathbb{P}[B]\mathbb{P}[C] = (1/2)(1/2)(4/36) = 1/36$ since there are four possibilities for the combinations of the rolls to get 9. But also $\mathbb{P}[A \cap B \cap C] = 1/36$ since we need a 3 in the first roll and thus a 6 in the second roll. These are the same probability so we have mutual independence. □

2. A and B are not pairwise independent.

   Solution. It's easy to see that the only element shared is 3, and so knowing that B happens puts a lot of information as to the outcome of A. Similarly, for the pairs of events (B, C) and (A, C), we don't have pairwise independence. □

3 Random Variables

We cover some issue of terminology. Independent trials of an experiment are a sequence of independent stages.

Bernoulli trials are independent trials where each stage has two possible outcomes.

Consider an n-toss sequence of coins which come up heads with probability ρ, and let $p(k)$ be the probability that k heads come up in the n-toss sequence. Then it's a known result that

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

Here we give a summary of counting results to make the discrete random variable probability distributions make more sense. Henceforth, we assume that every object is distinguishable.

- The number of permutations of n objects is $n! = \prod_{i=1}^{n} i$.
- The number of k-permutations of n objects is $n!/(n-k)!$.
- Binomial coefficients: The number of ways to choose k objects out of n objects is $\binom{n}{k} = n!/(k!(n-k)!)$.
- Multinomial coefficients: The number of ways to partition n objects in r groups, where n_i objects are in the ith groups and $\sum_{i=1}^{r} n_i = n$, is given by $\binom{n}{n_1,\ldots,n_r} = n!/(\prod_{i=1}^{r} n_i!)$.

Now, we can really start dealing with random variables.

Definition 26 (Random Variable). A **random variable** is a function taking points in the sample space to the real interval. Formally, a random variable is a function $X : \Omega \to \mathbb{R}$

In the future we may generalize this but for now we deal with ones with codomain $\mathbb{R}$.

Definition 27 (Discrete Random Variable). A **discrete random variable** is a random variable whose domain is composed of at most countably many elements. That is, $|\Omega| \leq |\mathbb{N}|$.

Definition 28 (Continuous Random Variable). A **continuous random variable** is a random variable whose domain is composed of at least uncountably many elements. That is, $|\Omega| \geq |\mathbb{R}|$.

Some key concepts about random variables are:

- A function (that takes real numbers to real numbers) of a random variable is another random variable.
- With each random variable, we can associate attributes such as the mean, variance, and so on.
- A random variable can be conditioned on an event or another random variable.
- There is a notion of independence of a random variable, on another random variable or just an event.

Definition 29 (Probability Function). For a random variable X, we associate the probability function $p_X(x)$, which assigns a probability to each value in the range of X. For a discrete random variable, the **probability mass function** (PMF) $\sum_x p_X(x) = 1$; for a continuous random variable, the **probability density function** (PDF) $\int_x p_X(x) = 1$. This function must be positive semidefinite (bounded below by zero) and for a discrete random variable must be bounded above by 1.

Note that for a continuous random variable X, we know that $p_X(x) \neq \mathbb{P}[X = x]$; indeed, we know that $\mathbb{P}[X = x] = 0$ since it's just the measure in an uncountable space of countably many points. In particular, $p_X(x) = \lim_{\epsilon \to 0} \mathbb{P}[x \leq X \leq x + \epsilon]/\epsilon$.

In particular, $\mathbb{P}[X \in B] = \sum_{x \in B} p_X(x)$ for a discrete random variable and $\mathbb{P}[X \in B] = \int_{x \in B} p_X(x)\,\mathrm{d}x$ for a continuous random variable. We are using the Riemann integral, instead of Riemann-Stieltjes or Lebesgue. As a consequence, $p_X(x)$ is piecewise-continuous.

Definition 30 (Cumulative Function). The cumulative distribution of a random variable X is defined as $P_X(x) = \mathbb{P}[X \leq x]$.

For a discrete distribution, the **cumulative mass function** is given by $P_X(x) = \sum_{i=-\infty}^{x} p_X(i)$. For a continuous distribution, the **cumulative density function** is given by $P_X(x) = \int_{-\infty}^{x} p_X(x')\,\mathrm{d}x'$, and $p_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} P_X(x)$.

Definition 31 (Expectation). The **expectation** of a random variable X with probability mass function $p_X(x)$ is given by

$$\mathbb{E}[X] = \sum_x x p_X(x) = \int_x x p_X(x)\,\mathrm{d}x$$

if we have $\sum_x |x| p_X(x) < \infty$ or $\int_x |x| p_X(x)\,\mathrm{d}x < \infty$ as appropriate.

Example 32. Consider a random variable X which takes in values of the form 2^k for integer k and has the PMF:

$$p_X\left(2^k\right) = 2^{-k}$$

$\mathbb{E}[X]$ does not exist.

Proof. We compute $\mathbb{E}[X]$:

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} 2^i p_X\left(2^i\right) = \sum_{i=0}^{\infty} 2^i 2^{-i} = \sum_{i=0}^{\infty} 1$$

This is clearly a divergent sum and so the expectation does not exist. □

Theorem 33 (Linearity of Expectation). For random variables $X_1, \ldots, X_n$ and real numbers $a_1, \ldots, a_n$, we have

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

Note that this doesn't depend on independence of the X_i.

Lemma 34 (Expectation of Function). The expectation of a function f of a random variable X with probability mass function $p_X(x)$ is given by

$$\mathbb{E}[f(X)] = \sum_x f(x) p_X(x) = \int_x f(x) p_X(x)\,\mathrm{d}x$$

if we have $\sum_x |f(x)| p_X(x) < \infty$ or $\int_x |f(x)| p_X(x)\,\mathrm{d}x < \infty$ as appropriate.

Definition 35 (Variance). The **variance** of a random variable X is given by

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

Note that $\mathrm{Var}[X]$ is positive semidefinite, that is, $\mathrm{Var}[X] \geq 0$ for any X.

Definition 36 (Standard Definition). The **standard deviation** of a random variable X is given by

$$\sigma_X = \sqrt{\mathrm{Var}[X]}$$

Fact 37. For a random variable X, if $\mathrm{Var}[X] = 0$, then $X = \mathbb{E}[X]$ almost always (with probability 1).

Proof. Write $\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$. Then either we have $\mathrm{Var}[X] = \sum_x (X - \mathbb{E}[X])^2 p_X(x)$ or $\mathrm{Var}[X] = \int_x (X - \mathbb{E}[X])^2 p_X(x)\,\mathrm{d}x$. In both of these cases, we need $X = \mathbb{E}[X]$ with probability 1, otherwise the sum or integral will introduce extra positive components into the variance, since the mapping $x \mapsto x^2$ is positive semidefinite. □

Fact 38. If $g(X) + aX + b$ and X is a random variable and we define $Y = g(X)$, then

$$\mathbb{E}[Y] = a\mathbb{E}[X] + b$$

and

$$\operatorname{Var}[Y] = a^2 \operatorname{Var}[X]$$

Proof. Assume Y is discrete, the continuous case goes in the same way. By linearity of expectation, we have

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[aX + b] \\ &= \mathbb{E}[aX] + b \\ &= a\mathbb{E}[X] + b\end{aligned}$$

On the topic of variance, we have

$$\begin{aligned}\operatorname{Var}[Y] &= \sum_x (ax + b - \mathbb{E}[aX + b])^2 p_X(x) \\ &= \sum_x (ax + b - a\mathbb{E}[X] - b)^2 p_X(x) \\ &= a^2 \sum_x (x - \mathbb{E}[X])^2 p_X(x) \\ &= a^2 \operatorname{Var}[X]\end{aligned}$$

as claimed. □

Fact 39. For a random variable X, we have $\operatorname{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$.

Proof. Assume X is discrete, the continuous case goes in the same way. We have

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

as claimed. □

Definition 40 (Joint Distribution). The **joint probability function** (mass or density) of two random variables is the function given by

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

where the latter two are equivalent to the first in the discrete case.

Note that this function still has to integrate over its domain to 1.

Definition 41 (Joint Cumulative Distribution). If X and Y are random variables, then $P_{X,Y}(x, y) = \mathbb{P}[X < x, Y < y]$.

If X and Y are continuous with joint density $P_{X,Y}(x, y)$, then $p_{X,Y}(x, y) = \frac{\partial^2 P_{X,Y}(x,y)}{\partial x \partial y}$.

Definition 42 (Marginal Distribution). The **marginal distribution** of a random variable X with respect to a random variable Y is just

$$p_X(x) = \sum_y p_{X,Y}(x, y) \text{ or } \int_y p_{X,Y}(x, y)\,\mathrm{d}y$$

depending on whether Y is discrete or continuous.

In particular, we have $p_X(x) = \sum_y p_{X,Y}(x, y)$ or $\int_y p(X, Y)(x, y)$ and $p_Y(y) = \sum_x p_{X,Y}(x, y)$ or $\int_x p_{X,Y}(x, y)\,\mathrm{d}x$.

Fact 43. As we noted, for a random variable X, $\mathbb{E}[g(X)] = \sum_x g(x)p_X(x)$ or $\int_x g(x)p_X(x)\,\mathrm{d}x$. Similarly,

$$\mathbb{E}[g(X,Y)] = \sum_x \sum_y g(x,y)p_{X,Y}(x,y) \text{ or } \int_x \int_y g(x,y)p_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x$$

Example 44 (Expectation of Multivariate Linear Function). Find the expectation $\mathbb{E}[aX+bY+c]$.

Solution. Write

$$\mathbb{E}[aX+bY+c] = \mathbb{E}[aX+bY] + c = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$$

by linearity. □

Definition 45 (Conditional Probability). If X is a random variable and A is an event, then

$$p_{X|A}(X = x \mid A) = \frac{\mathbb{P}[X = x \cap A]}{\mathbb{P}[A]}$$

Corollary 46. For a random variable X and event A, we know that

$$\sum_x p_{X|A}(x \mid A) = 1 \text{ or } \int_x p_{X|A}(x \mid A)\,\mathrm{d}x = 1$$

Definition 47 (Conditioning With Another Random Variable). Let X and Y be random variables. Then

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

Hence, $\sum_x p_{X|Y}(x \mid y) = 1$ or $\int_x p_{X|Y}(x \mid y)\,\mathrm{d}x = 1$.

The extension to multiple random variables is something like $p_{X,Y|Z}(x,y \mid z)$ or $p_{X|Y,Z}(x \mid y,z)$. Similar definitions can be used for these multiple random variables.

From this we can derive Bayes' rule for distributions. In particular,

Theorem 48 (Bayes' Rule on Distribution Functions). If X and Y are random variables with probability functions p_X and p_Y, we have

$$p_{X|Y}(x \mid y) = \frac{p_{Y|X}(y \mid x)p_X(x)}{p_Y(y)} = \frac{p_{Y|X}(y \mid x)p_X(x)}{\sum_{x'} p_{Y|X}(y \mid x')p_X(x')} \quad \text{or} \quad \frac{p_{Y|X}(y \mid x)p_X(x)}{\int_{x'} p_{Y|X}(y \mid x')p_X(x')\,\mathrm{d}x'}$$

Definition 49 (Conditional Expectation). Suppose A is an event where $\mathbb{P}[A] > 0$. Then

$$\mathbb{E}[X \mid A] = \sum_x x p_{X|A}(x \mid A) \text{ or } \int_x x p_{X|A}(x \mid A)\,\mathrm{d}x$$

In particular,

$$\mathbb{E}[X \mid Y = y] = \sum_x x p_{X|Y}(x \mid y) \text{ or } \int_x x p_{X|Y}(x \mid y)$$

Theorem 50 (Total Expectation Theorem). Let $A_1, \ldots, A_n$ be a disjoint events, where $\mathbb{P}[A_i] > 0$ for each i, that forms a partition of Ω. Then

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

Proof. We cover the discrete case, the continuous case goes the same way.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_x x p_X(x) \\
&= \sum_x x \sum_{i=1}^n \mathbb{P}[A_i] p_{X|A_i}(x) && \text{(Total Probability Theorem)}
\end{aligned}$$

$$= \sum_{i=1}^{n} \mathbb{P}[A_i] \sum_x x\mathbb{P}[X \mid A_i](x) \qquad \text{(Fubinu's Theorem in continuous case)}$$
$$= \sum_{i=1}^{n} \mathbb{P}[A_i]\mathbb{E}[X \mid A_i]$$

□

Mostly, since we use this for analysis of random variables, we use the form

$$\mathbb{E}[X] = \sum_y p_Y(y)\mathbb{E}[X \mid Y = y] \text{ or } \int_y p_Y(y)\mathbb{E}[X \mid Y = y]\,\mathrm{d}y$$

but this is just a restatement of the total expectation theorem.

Note that $\mathbb{E}[X \mid Y]$ can also be thought of as a random variable, and so can $X \mid Y$ itself. To see this, just look at the definition of a random variable.

Theorem 51 (Iterated Expectation). Continuing from the total expectation theorem, in the discrete case we have

$$\mathbb{E}[X] = \sum_y p_Y(y)\mathbb{E}[X \mid Y = y] = \mathbb{E}[\mathbb{E}[X \mid Y = y]] = \mathbb{E}[\mathbb{E}[X \mid Y]]$$

and in the continuous case we have

$$\mathbb{E}[Y] = \int_y p_Y(y)\mathbb{E}[X \mid Y = y] = \mathbb{E}[\mathbb{E}[X \mid Y = y]] = \mathbb{E}[\mathbb{E}[X \mid Y]]$$

Definition 52 (Independence of Random Variables). Let X be a random variable and A be an event. Then X is independent of A if and only if

$$\mathbb{P}[X = x \cap A] = p_X(x)\mathbb{P}[A] \text{ or } p_{X|A}(x \mid A) = p_X(x)\mathbb{P}[A]$$

Let Y be a random variable. Then X and Y are independent if and only if

$$p_{X,Y}(x, y) = p_X(x)p_Y(y)$$

for any x and y.

Corollary 53. If X and Y are random variables, and $p_Y(y) > 0$ for any y, then X and Y are independent if and only if

$$p_{X|Y}(x \mid y) = p_X(x) \quad \text{or} \quad p_{Y|X}(y \mid x)$$

Lemma 54. If X and Y are independent random variables, then

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

Proof. We solve the discrete case, the continuous case goes the same way. By definition,

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_x \sum_y xyp_{X,Y}(x, y) \\
&= \sum_x \sum_y xyp_X(x)p_Y(y) \\
&= \sum_x xp_X(x) \sum_y yp_Y(y) \\
&= \sum_x xp_X(x)\mathbb{E}[Y] \\
&= \mathbb{E}[Y] \sum_x xp_X(x) \\
&= \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}$$

as claimed. □

Definition 55 (Covariance). For random variables X and Y, the **covariance** of X and Y is given by

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

Corollary 56. For a random variable X,

$$\mathrm{Cov}[X,X] = \mathrm{Var}[X]$$

Claim 57. If X and Y are random variables, then

$$\mathrm{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Proof. By expansion:

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

as claimed. □

Definition 58 (Correlation). The **correlation** between random variables X and Y is given by

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

Corollary 59. We know that $\rho(X,Y) = 0$ if and only if $\mathrm{Cov}[X,Y] = 0$ and $\mathrm{Var}[X] > 0$ and $\mathrm{Var}[Y] > 0$. In this case we say that X and Y is **uncorrelated**.

The correlation coefficient $\rho(X,Y)$ measures how the random variables X and Y behave around the mean.

Note that $\mathrm{Cov}[X,Y] = 0$ does not imply that X and Y are independent.

Example 60. If the joint distribution of X and Y is given by $p_{X,Y}(x,y) = 0$ unless $(x,y) \in \{(0,-1),(0,1),(1,-1),(1,1)\}$, in which case it takes $1/4$, then $p_X(x) = 1/4$ if $x = -1$, $1/2$ if $x = 0$, and $1/4$ if $x = 1$, and the same distribution for $p_Y(y)$. In this case, $\mathbb{E}[XY] = 0$ since at least one of X or Y is zero for all points where the probability is nonzero. Also, X and Y are both symmetric around 0, $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ so $\mathrm{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$. However, X and Y are not independent, since giving a value for X determines at most two values for Y, whereas the original Y has four outcomes.

Most of the time, this is true, but in the special case of the (multivariate) normal distribution, we can use $\mathrm{Cov}[X,Y] = 0$ to show that X and Y are independent.

Claim 61. If X and Y are random variables, then $|\rho(X,Y)| \le 1$.

Proof. We begin with a small claim.

Claim 61.1. The *Schwarz inequality* on random variables states that

$$(\mathbb{E}[XY])^2 \le \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]$$

Proof. If we have $\mathbb{E}\left[Y^2\right] = 0$ then $\mathbb{P}[Y=0] = 1$, so $\mathbb{E}[XY] = 0$. If either X or Y are 0 with probability 1, then the inequality holds in this manner.

Now assume $\mathbb{E}\left[Y^2\right] > 0$ and $\mathbb{E}\left[X^2\right] > 0$. Then

$$0 \le \mathbb{E}\left[\left(X - Y\frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}\right)^2\right]$$

$$0 \le \mathbb{E}\left[X^2\right] - 2\frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]} + \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]^2}$$
$$\frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]} \le \mathbb{E}\left[X^2\right]$$
$$\mathbb{E}[XY]^2 \le \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]$$

as claimed. □

Now we continue with the main proof. Let $\widetilde{X} = X - \mathbb{E}[X]$ and $\widetilde{Y} = Y - \mathbb{E}[X]$. By the definition of the correlation coefficient, and the Schwarz inequality, we have

$$\rho(X,Y)^2 = \frac{\mathbb{E}\left[\widetilde{X}\widetilde{Y}\right]}{\mathbb{E}\left[\widetilde{X}\right]\mathbb{E}\left[\widetilde{Y}\right]} \le 1$$

□

Claim 62. If X and Y are independent random variables, then $\operatorname{Var}[X+Y] = \operatorname{Var}[X] + \operatorname{Var}[Y]$.

Proof. Define $\widetilde{X} = X - \mathbb{E}[X]$ and $\widetilde{Y} - \mathbb{E}[Y]$. These random variables have $\mathbb{E}\left[\widetilde{X}\right] = \mathbb{E}\left[\widetilde{Y}\right] = 0$.

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

In this way we can show inductively that for independent random variables $X_1, \dots, X_n$, we have $\operatorname{Var}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \operatorname{Var}[X_i]$. □

Definition 63 (Entropy). Let X be a random variable. Then the **entropy**, or average self-information, of X is defined as

$$H(X) = -\sum_x p_X(x)\log(p_X(x)) \text{ or } -\int_x p_X(x)\log(p_X(x))\,\mathrm{d}x$$

The term $-\log(p_X(x))$ is the self-information of the event $X = x$.

Example 64. The entropy of $X \sim \text{Bernoulli}(p)$ is maximal when $p = 1/2$.

Definition 65 (Order Statistics). The **k^{th} order statistic** refers to the k^{th} smallest random variables from a set of random variables $X_1, \dots, X_n$.

In particular, we are interested in the smallest and the largest random variables.

Suppose $X_1, \dots, X_n$ are i.i.d.random variables with cumulative density functions $P_X(x)$. Let $Y = \min(X_1, \dots, X_n)$, and $Z = \max(X_1, \dots, X_n)$.

Claim 66. The cumulative density $P_Y(y) = 1 - (1 - P_X(y))^n$. Also, $P_Z(z) = F_X(z)^n$.

Proof. By simple computation:

$$P_Y(y) = \mathbb{P}[Y \le y]$$

$$\begin{aligned}
&= \mathbb{P}[\min(X_1, \ldots, X_n) \le y] \\
&= 1 - \mathbb{P}[\min(X_1, \ldots, X_n)] \\
&= 1 - \prod_{i=1}^{n} \mathbb{P}[X_i \ge y] \\
&= 1 - (1 - P_X(y))^n
\end{aligned}$$

Also,

$$\begin{aligned}
P_Z(z) &= \mathbb{P}[Z \le 3] \\
&= \mathbb{P}[\max(X_1, \ldots, X_n) \le z] \\
&= \prod_{i=1}^{n} \mathbb{P}[X_i \le z] \\
&= P_X(z)^n
\end{aligned}$$

as desired. □

Definition 67 (Convolution). Let X and Y be independent random variables with probability functions $p_X(x)$ and $p_Y(y)$. Let $Z = X + Y$. Then

$$p_Z(z) = \sum_{(x,y)=z} p_{X+Y}(x+y) = \sum_x p_{X,Y}(x, z-x) = \sum_x p_X(x) p_Y(z-x)$$

in the discrete case, or

$$p_Z(z) = \int_{(x,y)=z} p_{X+y}(x+y)\,\mathrm{d}x\,\mathrm{d}y = \int_x p_{X,Y}(x, z-x)\,\mathrm{d}x = \int_x p_X(x) p_Y(z-x)\,\mathrm{d}x$$

in the continuous case, is the **convolution** $X * Y$ of X and Y, or equivalently $p_Z(z) = p_X(x) * p_Y(z-x)$.

Claim 68. The latter two definitions in the continuous case are equivalent.

Proof. We condition on the value of x:

$$\begin{aligned}
P_{Z|X}(z \mid x) &= P_{Z|X}(x+y \mid x) \\
&= P_{Y|X}(z-x \mid x) \\
&= P_Y(z-x) && \text{(Independence of } X \text{ and } Y.\text{)}
\end{aligned}$$

Differentiation obtains $p_{Z|X}(z \mid x) = p_Y(z-x)$. The multiplication rule gives us $p_{X,Z}(x,z) = p_X(x) p_{Z|X}(z \mid x)$, and finally, by independence again,

$$p_Z(z) = \int_x p_{X,Z}(x,z)\,\mathrm{d}x = \int_x p_X(x) p_Y(z-x)\,\mathrm{d}x$$

as claimed. □

Definition 69 (Moment Generating Function). For a random variable X, we define the **moment generating function** $\mathrm{MGF}(X) = M_X(s) = \mathbb{E}[\exp(sX)]$ for a real number s.

How does this generate the moments? Assume X is continuous. Then

$$M_X(s) = \int_x \exp(sx) p_X(x)\,\mathrm{d}x$$

Taking the derivative n times and using integration by parts and using Feynman trick (Leibniz integral rule), we have

$$\begin{aligned}
\frac{\mathrm{d}^n}{\mathrm{d}s^n} M_X(s) &= \frac{\mathrm{d}^n}{\mathrm{d}s^n} \int_x \exp(sx) p_X(x)\,\mathrm{d}x \\
&= \int_x \frac{\partial^n}{\partial s^n} \exp(sx) p_X(x)\,\mathrm{d}x
\end{aligned}$$

$$= \int_x x^n \exp(sx) p_X(x)\,\mathrm{d}x$$

Finally, plugging in $s = 0$ gets the nth moment of X.

Fact 70. The moment generating function $M_X(s)$ uniquely specifies both $p_X(x)$ and $P_X(x)$ of a random variable.

Fact 71. Let X and Y be independent random variables, and $Z = X + Y$. Then $M_Z(s) = M_X(s)M_Y(s)$.

Proof. By expansion:

$$\begin{aligned} M_Z(s) &= \mathbb{E}\Big[\exp(s(X+Y))\Big] \\ &= \mathbb{E}\Big[\exp(sx)\exp(sy)\Big] \\ &= M_X(s)M_Y(s) \end{aligned}$$

as desired. □

We can do this inductively to show that this is true for N independent random variables instead of just two.

Fact 72. If $X = \sum_{i=1}^n X_i$, then $M_X(s) = \prod_{i=1}^n M_{X_i}(s)$.

Fact 73. If $M_X(s)$ is finite over $(-a, a)$ for some $a > 0$, then the moment generating function uniquely determines the cumulative distribution function for the random variable X.

Example 74. If $M(s) = \frac{1}{4}\exp(-s) + \frac{1}{2} + \frac{1}{8}\exp(4s) + \frac{1}{8}\exp(5s)$, then $p_X(x) = 1/4$ if $x = 1$, $p_X(x) = 1/2$ if $x = 0$, $p_X(x) = 1/8$ if $x = 4$, and $p_X(x) = 1/8$ if $x = 5$. This is by definition of expectation.

Example 75. Suppose for a random variable X that $M_X(s) = p\exp(s)/(1 - (1-p)\exp(s))$. Then

$$M_X(s) = p\exp(s)\left(\sum_{i=0}^{\infty}((1-p)\exp(s))^i\right) = \sum_{i=0}^{\infty} p(1-p)^i \exp(s)^{i+1}$$

Then $p_X(k) = p(1-p)^{k-1}$, for all $k \geq 1$, assuming $(1-p)\exp(s) < 1$ or $s < \log(1/(1-p))$. To see this, differentiate by s, and obtain

$$\frac{\mathrm{d}}{\mathrm{d}s}M_X(s) = \frac{\mathrm{d}}{\mathrm{d}s}\sum_{i=0}^{\infty} p(1-p)^i \exp(s)^{i+1} = \sum_{i=0}^{\infty}(i+1)p(1-p)^i\exp(s)^{i+1} = \sum_{i=1}^{\infty} ip(1-p)^{i-1}\exp(s)^i$$

Plugging in $s = 0$, we get

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} ip(1-p)^{i-1}$$

which gives, by definition, that $p_X(x) = p(1-p)^{x-1}$.

Theorem 76 (Markov Inequality). If a random variable X is positive semidefinite, then for every $a > 0$

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

Proof. Fix an $a > 0$. Define $Y = a \cdot \mathbb{1}(X \geq a)$. Then since $Y \leq X$, we know $\mathbb{E}[Y] \leq \mathbb{E}[X]$. Then $\mathbb{E}[Y] = 0 \cdot p_Y(0) + a \cdot p_Y(a) = a \cdot p_Y(a) = a\mathbb{P}[X \geq a]$. Thus $a\mathbb{P}[X \geq a] \leq \mathbb{E}[X]$, so $\mathbb{P}[X \geq a] \leq \mathbb{E}[X]/a$, as desired. □

Theorem 77 (Chebyshev Inequality). If X is a random variable with finite mean and variance, then for every $c > 0$

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

Proof. Fix a $c > 0$. Then

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq c] = \mathbb{P}\left[(X - \mathbb{E}[X])^2 \geq c^2\right] \leq \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{c^2}$$

where the last step is by Markov's inequality. □

Corollary 78. If X is a random variable with finite mean and variance, then for each $k > 0$

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq k\sqrt{\text{Var}[X]}\right] \leq \frac{1}{k^2}$$

Proof. Plug in $c = k\sqrt{\text{Var}[X]}$ into Chebyshev's inequality. □

Example 79. Consider $X \sim \text{Normal}\left(0, \sigma^2\right)$. Then to determine the level of confidence of an interesting result X a priori hypothesized to have mean μ, we compute $1 - \mathbb{P}[|X - \mu| \leq k\sigma] \leq \frac{1}{k^2}$. In particular, $\mathbb{P}[|x - \mu| \geq 2\sigma] \leq \frac{1}{4}$.

Theorem 80 (Chernoff Bound). Let $M_X(s)$ be the moment generating function of a random variable X. In particular, $M_X(s) = \mathbb{E}[\exp(sX)]$. For every a and every $s \geq 0$, then

$$\mathbb{P}[X \geq a] \leq \exp(-sa)M_X(s)$$

Proof. Fix an $s > 0$. Then

$$\mathbb{P}[X \geq a] = \mathbb{P}[sX \geq sa] = \mathbb{P}\left[\exp(sX) \geq \exp(sa)\right] \leq \frac{\mathbb{E}[sX]}{\exp(sa)} = \exp(-sa)M_X(s)$$

by the Markov inequality.

If $s = 0$, then it's obviously true. □

Corollary 81. For every a and every $s \leq 0$,

$$\mathbb{P}[X \leq a] \leq \exp(sa)M_X(s)$$

Proof. Suppose $s < 0$. Then

$$\mathbb{P}[X \leq a] = \mathbb{P}[sX \leq sA] = \mathbb{P}\left[\exp(sX) \leq \exp(sa)\right] \leq \frac{\mathbb{E}[\exp(sX)]}{\exp(sa)} = \exp(sa)M_X(s)$$

If $s = 0$, then it's obviously true. □

Note that we can tighten the bound by minimizing $M_X(s)$. In particular, we have the following corollary:

Corollary 82. For every $a \geq 0$, we have

$$\mathbb{P}[X \geq a] \leq \exp(-\phi(a))$$

where $\phi(a) = \max_{s \geq 0}\left(sa - \log(M_X(s))\right)$.

Proof. By Chernoff,

$$\begin{aligned}
\mathbb{P}[X \geq a] &\leq \min_{s \geq 0}\left(\exp(-sa)M_X(s)\right) \\
&= \min_{s \geq 0}\exp\left(-sa - \log(M_X(s))\right) && \text{(Exponential map is monotone)} \\
&= \exp\left(-\max_{s \geq 0}(sa - \log(M_X(s)))\right) \\
&= \exp(-\phi(a))
\end{aligned}$$

□

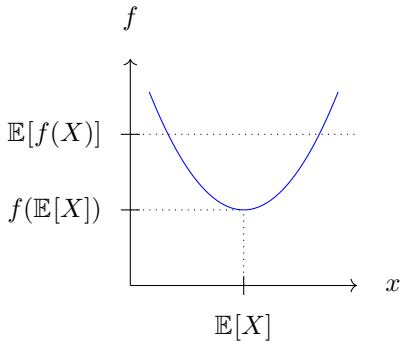
This is very useful for finding how rare events occur in random processes, in a field called "Large Deviations Theory".

Theorem 83 (Jensen's Inequality). Let f be a twice differentiable convex function (which means that $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2} \geq 0$ for all x). Then

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

Proof. Long and boring. □

An intuitionist perspective:



Discrete Random Variables

With each discrete random variable, we associate a **probability mass function**.

Example 84. Consider a pair of four-sided die, and consider the Ω that is the sample space of the values of the rolls of the dice. Then consider $X(\omega) = \max(\text{the two rolls' values})$. This is a legitimate discrete random variable. What is the probability mass function of X?

Solution. The probability mass function $\mathrm{PMF}(X)$ is $\mathbb{P}[X = 1] = 1/16$, $\mathbb{P}[X = 2] = 3/16$, $\mathbb{P}[X = 3] = 5/16$, and $\mathbb{P}[X = 4] = 9/16$. □

In general, to construct a probability mass function, for each $k \in \mathrm{range}(X)$, take the subset of points $S \subseteq \Omega$ which map to k, and compute $|S|/|\Omega|$.

Definition 85 (Bernoulli Random Variable). The **Bernoulli Random Variable** is a random variable which takes a parameter $p \in [0, 1]$. If X is Bernoulli with parameter p, we say $X \sim \mathrm{Bernoulli}(p)$. The probability mass function of X is given by

$$\mathrm{PMF}(X) = p_X(k) = kp + (1 - p)(1 - k)$$

or the equivalent case definition:

$$p_X(k) = \begin{cases} p & k = 1 \\ 1 - p & k = 0 \end{cases}$$

The Bernoulli random variable is used, for example, to compute the outcome of a single coin toss.

Fact 86. If $X \sim \mathrm{Bernoulli}(p)$, then $\mathbb{E}[X] = p$ and $\mathrm{Var}[X] = p(1 - p)$.

Proof. For the expectation, we have $\mathbb{E}[X] = p \cdot 1 + (1 - p) \cdot 0 = p$, and for the variance, $\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = p - p^2 = p(1 - p)$ as claimed. □

Fact 87. IF $X \sim \mathrm{Bernoulli}(p)$, then $M_X(s) = 1 - p + p\exp(s)$.

Definition 88 (Binomial Random Variable). The **Binomial Random Variable** is a random variable which takes two parameters $n \in \mathbb{N}$ and $p \in [0,1]$. If X is binomial with parameters n and p, we say $X \sim \text{Binomial}(n,p)$. The probability mass function of X is given by

$$\text{PMF}(X) = p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The binomial random variable is used to compute, for example, the number of heads in n coin tosses.

Note that, for large n and reasonable p, the binomial distribution can be approximated (looks like) a normal distribution. We'll go over the central limit theorem, which tells us why this is true.

Corollary 89 (Binomial Theorem). The **Binomial Theorem** is the following identity: for any $n \in \mathbb{N}$ and $p \in [0,1]$ we have

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$$

Fact 90. If $X_1, \dots, X_n \sim \text{Bernoulli}(p)$ are **independent and identically distributed** (i.i.d.), then if $Y = \sum_{i=1}^n X_I$, then $Y \sim \text{Binomial}(n,p)$.

Fact 91. If $X \sim \text{Binomial}(n,p)$, then $\mathbb{E}[X] = np$ and $\text{Var}[X] = np(1-p)$.

Proof. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. Then $X = \sum_{i=1}^n X_i$. So

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

Also,

$$\text{Var}[X] = \text{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \text{Var}[X_i] = np(1-p)$$

as claimed. □

Fact 92. If $X \sim \text{Binomial}(n,p)$, then $M_X(s) = (1 - p + p\exp(s))^n$.

Proof. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, and $X = \sum_{i=1}^n X_i$. Then

$$M_X(s) = \prod_{i=1}^n M_{X_i}(s) = \prod_{i=1}^n \left(1 - p + p\exp(s)\right) = (1 - p + p\exp(s))^n$$

as claimed. □

Definition 93 (Geometric Random Variable). The **Geometric Random Variable** is a random variable which takes one parameter $p \in [0,1]$. If X is geometric with parameter p, we say $X \sim \text{Geometric}(p)$. The probability mass function of X is given by

$$\text{PMF}(X) = p_X(k) = (1-p)^{k-1} p$$

where the $(1-p)^{k-1}$ term indicates $k-1$ failures and the p term indicates 1 success, with no freedom on how to place the success (at the end).

The geometric random variable can be used to compute, for example, the number of tosses until we get a head.

Fact 94. The geometric distribution is a legitimate distribution. We have

$$\sum_{k=1}^{\infty} (1-p)^{k-1} p = p \sum_{k=0}^{\infty} (1-p)^k = \frac{p}{1-(1-p)} = 1$$

so we're good.

Fact 95. If $X \sim \text{Geometric}(p)$, then $\mathbb{E}[X] = p^{-1}$ and $\text{Var}[X] = (1-p)^{p-2}$.

Proof. For the expectation, we have

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=1}^{\infty} k(1-p)^k p \\ &= -p\sum_{k=0}^{\infty} \frac{\mathrm{d}}{\mathrm{d}p}(1-p)^k \\ &= -p\frac{\mathrm{d}}{\mathrm{d}p}\sum_{k=0}^{\infty}(1-p)^k \\ &= -p\frac{\mathrm{d}}{\mathrm{d}p}\left(\frac{1}{1-(1-p)}\right) \\ &= -p\frac{\mathrm{d}}{\mathrm{d}p}\frac{1}{p} \\ &= \frac{1}{p}\end{aligned}$$

Similarly, $\mathbb{E}\left[X^2\right] = 2p^{-2} - p^{-1}$ (by differentiating twice). Hence, $\text{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = 2p^{-2} - p^{-1} - p^{-2} = (1-p)p^{-2}$. □

Fact 96. The geometric distribution is **memoryless**; that is,

$$\mathbb{P}[X = a + b \mid X \geq a] = \mathbb{P}[X = b]$$

Corollary 97. By the total expectation theorem, if $X \sim \text{Geometric}(p)$, then

$$\mathbb{E}[X] = \mathbb{P}[X=1]\mathbb{E}[X \mid X = 1] + \mathbb{P}[X > 1]\mathbb{P}[X \mid X > 1] = p \cdot 1 + (1-p)\cdot(1 + \mathbb{E}[X])$$

Similarly,

$$\mathbb{E}\left[X^2\right] p \cdot 1 + (1-p)\mathbb{E}\left[(1+X)^2\right]$$

which gives the results we already derived.

Definition 98 (Poisson Random Variable). The **Poisson Random Variable** is a random variable which takes one parameter $\lambda \in \mathbb{R}_{>0}$. If X is Poisson with parameter λ, we say $X \sim \text{Poisson}(\lambda)$. The probability mass function of X is given by

$$\text{PMF}(X) = p_X(k) = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

The Poisson random variable is essentially a sum of Bernoulli random variables which are independent but not necessarily identically distributed. If the ith Bernoulli random variable has parameter ip_i and the sequence $ip_i \to \lambda$ as $i \to \infty$, then we have a Poisson random variable in the limit.

Fact 99. The Poisson distribution is a legitimate distribution. We have

$$\sum_{k=0}^{\infty} \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!} = \mathrm{e}^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = 1$$

Fact 100. If $X \sim \text{Poisson}(\lambda)$, then $\mathbb{E}[X] = \lambda$ and $\text{Var}[X] = \lambda$.

Proof. For the expectation, we have

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k\frac{\mathrm{e}^{-\lambda}x^{-k}}{k!}$$

$$
\begin{aligned}
&= \lambda \mathrm{e}^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda \mathrm{e}^{-\lambda} \sum_{k=0}^{\infty} \frac{x^k}{k!} \\
&= \lambda \mathrm{e}^{-\lambda} \mathrm{e}^{\lambda} \\
&= \lambda
\end{aligned}
$$

as claimed. For the second moment, we have

$$
\begin{aligned}
\mathbb{E}\left[X^2\right] &= \sum_{k=0}^{\infty} k^2 \frac{\mathrm{e}^{-\lambda}\lambda^2}{k!} \\
&= \left(\sum_{k=0}^{\infty} k(k-1) \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!} \right) + \left(\sum_{k=0}^{\infty} k \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!} \right) \\
&= \lambda^2 \mathrm{e}^{-\lambda} \left(\sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} \right) + \lambda \\
&= \lambda^2 \mathrm{e}^{-\lambda} \mathrm{e}^{\lambda} + \lambda \\
&= \lambda^2 + \lambda
\end{aligned}
$$

Hence, $\operatorname{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \lambda$. □

Fact 101. If $X \sim \text{Poisson}(\lambda)$, then $M_X(s) = \exp(\lambda(\exp(s) - 1))$.

Proof. We know that $p_X(x) = \exp(-\lambda)\lambda^x / x!$. Then

$$
\begin{aligned}
M_X(s) &= \sum_{x=0}^{\infty} \exp(sx) \frac{\lambda^x \exp(-\lambda)}{x!} \\
&= \exp(-\lambda) \sum_{x=0}^{\infty} \frac{a^x}{x!} \qquad \text{where } a = \lambda \exp(s) \\
&= \exp\Big(\lambda \Big(\exp(s) - 1 \Big) \Big)
\end{aligned}
$$

as desired. □

Fact 102. If X and Y are independent Poisson random variables with parameters λ and μ, and $Z = X + Y$, then Z is a Poisson random variable with parameter $\lambda + \mu$.

Proof. Take the moment generating function:

$$
\begin{aligned}
M_Z(s) &= M_X(s) M_Y(s) \\
&= \exp\Big(\lambda(\exp(s) - 1) \Big) \exp\Big(\mu(\exp(s) - 1) \Big) \\
&= \exp\Big((\lambda + \mu)(\exp(s) - 1) \Big)
\end{aligned}
$$

Since the moment generating function uniquely determines $p_X(x)$ and $P_X(x)$, we know that $Z \sim \text{Poisson}(\mu + \lambda)$ as claimed. □

Definition 103 (Uniform Random Variable). The **Uniform Random Variable** is a random variable which takes two parameters $a, b \in \mathbb{R}$ with $b > a$. If X is uniformly distributed with parameters a and b, we say $X \sim$

Uniform$([a,b])$. The probability mass function of X is given by

$$\text{PMF}(X) = p_X(k) = \begin{cases} \frac{1}{b-a+1} & k \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

Continuous Random Variables

Example 104. Assume we uniformly sample a point from the unit ball. Then $\mu(O) = 0$, $f_X(0) = 1/V(B_2) = 1/(4\pi/3) = 3/(4\pi)$, $\mu(\partial B_2) = 0$ (the surface of simple manifold object has measure 0), and then $\mu(|X| < 1/2) = (4\pi(1/2)^3/3)/(4\pi/3) = (1/2)^3 = 1/8$.

Example 105. Assume we uniformly sample a point from the unit ball, and let R be its distance from the origin. Then $P_R(r) = (3/(4\pi))(4\pi r^3/3) = r^3$. As a result, $p_R(r) = \frac{\mathrm{d}}{\mathrm{d}r}P_R(r) = \frac{\mathrm{d}}{\mathrm{d}r}r^3 = 3r^2$. Then $\mathbb{E}[R] = \int_0^1 r \cdot p_R(r)\,\mathrm{d}r = \int_0^1 r \cdot 3r^2\,\mathrm{d}r = 3\int_0^1 r^3\,\mathrm{d}r = 3/4$.

Definition 106 (Continuous Uniform Random Variable). The **Continuous Uniform Random Variable** is a random variable which takes two parameters $a, b \in \mathbb{R}$ with $b > a$. If X is uniformly distributed with parameters a and b, we say $X \sim \text{Uniform}([a,b])$. The probability density function is given by

$$\text{PDF}(X) = p_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

Fact 107. If $X \sim \text{Uniform}([a,b])$, then $\mathbb{E}[X] = (a+b)/2$, and $\text{Var}[X] = (b-a)^2/12$.

Proof. We obtain the expectation by integrating:

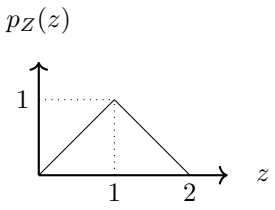
$$\begin{aligned} \mathbb{E}[X] &= \int_a^b \frac{x}{b-a}\,\mathrm{d}x \\ &= \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{a+b}{2} \end{aligned}$$

We also obtain the variance by integrating. Note that $\text{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$. Hence we have

$$\begin{aligned} \text{Var}[X] &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\ &= \left(\int_a^b \frac{x^2}{b-a}\,\mathrm{d}x\right) - \left(\frac{a+b}{4}\right)^2 \\ &= \left[\frac{x^3}{3(b-a)}\right]_a^b - \frac{(a+b)^2}{4} \\ &= \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

as desired. □

Example 108. If $X, Y \overset{\text{i.i.d.}}{\sim} \text{Uniform}([0,1])$, and $Z = X + Y$, then $f_Z(z) = \int_x f_X(x) f_Y(z-x)\,\mathrm{d}x$. This product is nonzero when $0 \le x \le 1$ and $0 \le z - x \le 1$, so $\max(0, z-1) \le x \le \min(1, z)$. This leads to four separate integrals. For $0 \le z \le 1$, we have $0 \le x \le z$; for $1 \le x \le 2$, we have $z - 1 \le x \le 1$. So $p_Z(z) = (\min(1,z) - \max(0, z-1)) \cdot \mathbb{1}(0 \le z \le 2)$. It looks like this:



Definition 109 (Exponential Random Variable). The **Exponential Random Variable** is a random variable which takes one parameter $\lambda > 0$. If X is exponentially distributed with parameter λ, we say $X \sim \text{Exponential}(\lambda)$. The probability density function is given by

$$\text{PDF}(X) = p_X(x) = \lambda \exp(-\lambda x)$$

Fact 110. If $X \sim \text{Exponential}(\lambda)$, then $\mathbb{E}[X] = 1/\lambda$ and $\text{Var}[X] = 1/\lambda^2$.

Proof. We obtain the expectation by integrating:

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty x\lambda \exp(-\lambda x)\,\mathrm{d}x \\ &= \left[-\frac{(1+\lambda x)\exp(-\lambda x)}{\lambda}\right]_0^\infty \\ &= \frac{1}{\lambda}\end{aligned}$$

We obtain the variance by integrating. We know that $\text{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$. So we have

$$\begin{aligned}\text{Var}[X] &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\ &= \left(\int_0^\infty x^2 \exp(-\lambda x)\,\mathrm{d}x\right) - \left(\frac{1}{\lambda}\right)^2 \\ &= \left[-\frac{\left(\lambda^2 x^2 + 2\lambda x + 2\right)\exp(-\lambda x)}{\lambda}\right]_0^\infty - \frac{1}{\lambda^2} \\ &= \frac{2}{k^2} - \frac{1}{k^2} \\ &= \frac{1}{k^2}\end{aligned}$$

as desired. □

Fact 111. If $X \sim \text{Exponential}(\lambda)$, then $P_X(x) = 1 - \exp(-\lambda x)$.

Proof. Integrate:

$$\begin{aligned}P_X(x) &= \int_{-\infty}^x p_X(x')\,\mathrm{d}x' \\ &= \int_{-\infty}^x \lambda \exp(-\lambda x')\,\mathrm{d}x' \\ &= 1 - \exp(-\lambda x)\end{aligned}$$

as claimed. □

Fact 112 (Exponential Memorylessness). The exponential distribution is **memoryless**: that is, if X is exponential, then

$$\mathbb{P}[X > a + b \mid X > a] = \mathbb{P}[X > b]$$

The exponential and geometric distributions are the only continuous and discrete random variables, respectively, which have this property.

Fact 113. In fact, the geometric distribution approaches the exponential distribution as the number of trials per unit time approaches infinity.

Proof. Let $X \sim \text{Geometric}(p)$ and $Y \sim \text{Exponential}(\lambda)$. Then $P_X(n) = 1 - (1-p)^n$. Define $\delta = -\log(1-p)/\lambda$ (which represents the number of trials per unit time), we have $\exp(-\lambda\delta) = 1 - p$. Thus, $P_X(n) = P_Y(n\delta)$. Taking $\lim_{\delta \to 0}$ while holding $n\delta$ constant, we interpret this as a geometric random variable holding a large amount of trials per unit time while making sure that the expected number of trials passed stays the same. □

Fact 114. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exponential}(\lambda)$. Let $Y = \min(X_1, \dots, X_n)$. Then $P_Y(y) = 1 - \exp(-n\lambda y)$.

Proof. We have $P_{X_i}(x) = 1 - \exp(-\lambda x)$. Then $P_Y(y) = 1 - (1 - (1 - \exp(-\lambda y)))^n = 1 - \exp(-n\lambda y)$. □

This is important because we deal with situations where this context appears when working on continuous time Markov chains.

Fact 115. If $X_1, \dots, X_n$ are independent and $X_i \sim \text{Exponential}(\lambda_i)$ for each i, and $Y = \min(X_1, \dots, X_n)$, then $Y \sim \text{Exponential}(\sum_{i=1}^n \lambda_i)$.

Fact 116. If $X \sim \text{Exponential}(\lambda)$, then $M_X(s) = \lambda/(\lambda - s)$ for $s < \lambda$.

Proof. We know that $p_X(x) = \lambda \exp(-\lambda x)$ for $x \geq 0$. We have

$$\begin{aligned} M_X(s) &= \lambda \int_0^\infty \exp(sx) \exp(-\lambda x) \, \mathrm{d}x \\ &= \lambda \int_0^\infty \exp(x(s-\lambda)) \, \mathrm{d}x \\ &= \lambda \left[\frac{\exp(x(s-\lambda))}{s-\lambda} \right]_0^\infty = \frac{\lambda}{\lambda + s} \end{aligned}$$

but this only works if $s < \lambda$, since otherwise the integral blows up. □

Definition 117 (Gaussian Random Variable). The **Gaussian (Normal) Random Variable** is a random variable which takes two real parameters μ (mean) and σ^2 (variance). If X is normally distributed with parameters μ and σ^2, we say $X \sim \text{Normal}(\mu, \sigma^2)$. The probability density function is

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

The Gaussian cumulative density function is usually given by $P_X(x) = \Phi(x)$, since it cannot be expressed by independent functions.

Fact 118. The sum of two independent Gaussians is Gaussian. If $X \sim \text{Normal}(\mu_1, \sigma_1^2)$ and $Y \sim \text{Normal}(\mu_2, \sigma_2^2)$ and $Z = X + Y$, then $Z \sim \text{Normal}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

However, the sum of two dependent Gaussians isn't always Gaussian; consider Gaussian X and $Y = X$ with probability $1/2$ and $Y = -X$ with probability $-1/2$. They're both Gaussian but $X + Y$ is not Gaussian. However, if $\mathbb{E}[X] \neq 0$, then Y is not a Gaussian.

Fact 119. A Gaussian multiplied by a constant is Gaussian. If $X \sim \text{Normal}(\mu, \sigma^2)$ and $Y = aX$, then $Y \sim \text{Normal}(a\mu, a^2\sigma^2)$.

These properties allow us to convert any Gaussian into the standard Gaussian.

Fact 120. If $X \sim \text{Normal}(\mu, \sigma^2)$, then if $Z = (X - \mu)/\sigma$ then $Z \sim \text{Normal}(0, 1)$.

Proof. Follows from the previous two results. □

Fact 121. If $X \sim \text{Normal}\big(\mu_X, \sigma_X^2\big)$ and $Y \sim \text{Normal}\big(\mu_Y, \sigma_Y^2\big)$, then if $Z = X + Y$ then

$$Z \sim \text{Normal}\big(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2\big)$$

Fact 122. If $Y \sim \text{Normal}(0, 1)$, then $M_y(s) = \exp\big(s^2/2\big)$.

Proof. We know that $p_Y(y) = \frac{1}{\sqrt{2\pi}} \exp\big(-y^2/2\big)$. Then

$$\begin{aligned} M_Y(s) &= \int_{-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{y^2}{2}\right) \exp(sy)\,\mathrm{d}y \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{2} + sy\right) \\ &= \frac{\exp\big(s^2/2\big)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{2} + sy - \frac{s^2}{2}\right) \mathrm{d}y \\ &= \frac{\exp\big(s^2/2\big)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(y-s)^2}{2}\right) \mathrm{d}y \\ &= \exp\left(\frac{s^2}{2}\right) \end{aligned}$$

since the final integral is just the integral of a normal distribution over its domain. □

Theorem 123 (Change of Variables). Let X be a random variable, and let $Y = f(X)$. Then $p_Y(y) = p_X(x) \cdot \left|\frac{\mathrm{d}f^{-1}(y)}{\mathrm{d}y}\right|$. More generally, if f is not invertible, let $f^{-1}(y)$ be the preimage of y under f. Then $P_Y(y) = P_X(f^{-1}(y)) = \sum_{z \in f^{-1}(y)} P_X(z)$.

Asymptotics and Convergence

Theorem 124 (Monotone Convergence Theorem). If $\{a_i\}$ is a monotone sequence of real numbers, then this sequence has a finite limit (i.e. $-\infty < \lim_{n\to\infty} a_n < \infty$) if and only if the sequence is bounded (i.e. there exists M such that $|a_i| < M$ for any i).

Proof. We begin with two lemmas:

Lemma 124.1. If a sequence of real numbers is increasing and bounded above, then its supremum is the limit.

Proof. Let $\{a_n\}$ be such a sequence. By assumption, $\{a_n\}$ is non-empty and bounded above. By the least-upper-bound property of real numbers, $c = \sup_n \{a_n\}$ exists and is finite. Now, for every $\epsilon > 0$, there exists N such that $a_N > c - \epsilon$, since otherwise $c - \epsilon$ is an upper bound of $\{a_n\}$, which contradicts to the definition of c. Then since $\{a_n\}$ is increasing, and c is its upper bound, for every $n > N$, we have $|c - a_n| \le |c - a_N| < \epsilon$. Hence, by definition, the limit of $\{a_n\}$ is $\sup_n \{a_n\}$. □

Lemma 124.2. If a sequence of real numbers is decreasing and bounded above, then its infimum is the limit.

Proof. Same as the above lemma. □

For the 'if' direction of the theorem, it follows directly from the lemmas. For the 'only if' direction, the definition of a limit implies any sequence with a limit is necessarily bounded. □

Definition 125 (Convergence in Probability). Let $X_1, \dots, X_n$ be random variables (not necessarily i.i.d.). We say that X_n converges to X in probability if for any $\epsilon > 0$ we have

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

We write $X_n \overset{\text{prob.}}{\to} X$.

Theorem 126 (Weak Law of Law Numbers). Let $X_1, \dots, X_n$ be i.i.d. random variables with mean μ and variance σ^2. Let $M_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then $M_n \overset{\text{prob.}}{\to} \mu$.

Proof. For $\mathbb{E}[M_n]$, we obtain $\mathbb{E}[M_n] = \mu$ by linearity. For $\text{Var}[M_n]$, we obtain $\text{Var}[M_n] = \sigma^2$ by independence of the X_i. Then by Chebyshev's inequality,

$$\mathbb{P}[|M_n - \mu| \geq \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}$$

and taking the limit gives that for any $\epsilon > 0$, we have that

$$\lim_{n\to\infty} \mathbb{P}[|M_n - \mu| \geq \epsilon] = 0$$

so $M_n \overset{\text{prob.}}{\to} \mu$. □

Example 127. Suppose that a fraction p of voters support a candidate. Let M_n be the fraction of voters sampled who support the candidate. Then

$$\mathbb{P}[|M_n - p| \geq \epsilon] = \frac{p(1-p)}{n\epsilon^2}$$

Since $\max_{0\leq p\leq 1} p(1-p) = 1/4$, so

$$\mathbb{P}[|M_n - p| \geq \epsilon] \leq \frac{1}{4n\epsilon^2}$$

To have the bound on the confidence level at $< \alpha$, we must have

$$\frac{1}{4n\epsilon^2} < \alpha \to n > \frac{1}{4\alpha\epsilon^2}$$

Definition 128 (Almost Sure Convergence). Let $X_1, \dots, X_n$ be a sequence of random variables, not necessarily i.i.d.. We say X_n converges almost surely to X, or with probability 1, if

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

We write $X_n \overset{\text{a.s.}}{\to} X$.

Theorem 129 (Strong Law of Large Numbers). Let $X_1, \dots, X_n$ be i.i.d. random variables with mean μ. Let $M_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then $M_n \overset{\text{a.s.}}{\to} \mu$, or

$$\mathbb{P}\left[\lim_{n\to\infty} M_n = \mu\right] = 1$$

Example 130. let $X_1, \dots, X_n$ be i.i.d. random variables uniformly distributed over $[0, 1]$. Let $Y_n = \min_i(X_i)$. Then Y_n are non-increasing, and since it's always non-negative, we have that Y_n must converge to some Y.

$$\mathbb{P}[Y_n \geq \epsilon] = \prod_{i=1}^n \mathbb{P}[X_i \geq \epsilon] = (1-\epsilon)^n$$

Thus,

$$\mathbb{P}[Y \geq \epsilon] = \mathbb{P}\left[\lim_{n\to\infty} Y_n \geq \epsilon\right] = \lim_{n\to\infty} (1-\epsilon)^n = 0$$

for any $\epsilon > 0$.

So $\mathbb{P}[\lim_{n\to\infty} Y_n = 0] = \mathbb{P}[Y = 0] = 1$, so $Y \overset{\text{a.s.}}{\to} 0$.

Definition 131 (Weak/Pointwise Convergence). A sequence of random variables X_n is said to converge to X in distribution, or converge weakly to X, if for every $\epsilon > 0$ and for every $a \in \mathbb{R}$ where $P_X(a)$ is continuous,

$$\lim_{n\to\infty} P_{X_n}(a) = P_X(a)$$

In this case, we say that P_{X_n} is pointwise convergent to P_X. Pointwise convergence can be applied to any sequence of functions that are discontinuous at at most finitely many points, not just cumulative distribution functions. Also, we write $X_n \overset{\text{weak}}{\to} X$, or $P_{X_n} \overset{\text{pointwise}}{\to} P_X$.

Theorem 132 (Central Limit Theorem). Let $X_1, \dots, X_n$ be i.i.d. random variables with mean μ and variance σ^2. Let $S_n = \sum_{i=1}^n X_i$, and define $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$. Then Z_n converges to the standard normal random variable in distribution, or $Z_n \overset{\text{weak}}{\to} \text{Normal}(0,1)$.

Note that $\mathbb{E}[Z_n] = 0$ and $\text{Var}[Z_n] = (n\sigma^2)/(n\sigma^2) = 1$.

Definition 133 (rth Moment Convergence). Let $X_1, \dots, X_n$ be random variables with finite kth moments for $0 \le k \le n$. Then X_n converges in the rth moment to X if

$$\lim_{n\to\infty} \mathbb{E}[(X_n - X)^r] = 0$$

for some $r > 0$.

We write $X_n \overset{L^r}{\to} X$.

Each type of convergence is related.

$$\begin{array}{ccccc} L^s \text{ convergence} & \xrightarrow{s \ge r \ge 1} & L^r \text{ convergence} & & \\ & & \downarrow & & \\ \text{almost sure convergence} & \longrightarrow & \text{convergence in probability} & \longrightarrow & \text{convergence in distribution} \end{array}$$

Or,

$$\begin{array}{ccccc} \mathbb{E}[(X - X_n)^s] \overset{n\to\infty}{\to} 0 & \xrightarrow{s \ge r \ge 1} & \mathbb{E}[(X - X_n)^r] \overset{n\to\infty}{\to} 0 & & \\ & & \downarrow & & \\ \mathbb{P}\left[\lim_{n\to\infty} X_n = X\right] = 1 & \longrightarrow & \forall \varepsilon > 0\ \mathbb{P}[|X_n - X| \ge \epsilon] \overset{n\to\infty}{\to} 0 & \longrightarrow & \forall \epsilon > 0\ \forall a \in \mathbb{R}\ P_{X_n}(a) \overset{n\to\infty}{\to} P_X(a) \end{array}$$

Example 134. Let $X \sim \text{Bernoulli}(p)$, for $0 < p < 1$, and $X_n = \left(1 + \frac{1}{n}\right)X$. Then $X_n - X = \frac{X}{n}$. Thus for any $\epsilon > 0$ $\mathbb{P}[\lim_{n\to\infty} |X_n - X| \ge \epsilon] = \mathbb{P}\left[\lim_{n\to\infty} \frac{X}{n} \ge x\right] = 0$, so by the complement event, $\mathbb{P}[\lim_{n\to\infty} |X_n - X| < \epsilon] = 1$. Thus, $X_n \overset{\text{a.s.}}{\to} X$ (and thus $X_n \overset{\text{prob.}}{\to} X$).

Corollary 135. If $\{X_n\}$ is a sequence of variables such that $\{X_n\} \overset{\text{a.s.}}{\to} X$, then $\{X_n\} \overset{\text{prob.}}{\to} X$.

Proof As Presented in Class. If $\{X_n\} \overset{\text{a.s.}}{\to} X$, then $\mathbb{P}[\lim_{n\to\infty} X_n = X] = 1$. Then, by the definition of limit, for any $\epsilon > 0$ we have there exists some function $m\colon \omega \to \mathbb{N}$ such that $\mathbb{P}[\lim_{n\to\infty} X_n = X] = \mathbb{P}[\omega\colon \lim_{\to\infty} X_n(\omega) = X(\omega)] = \mathbb{P}[\omega\colon \omega \in \Omega, |X_n(\omega) - X(\omega)| < \epsilon\ \forall n > m(\omega)] = 1$.

Fix an $\epsilon > 0$, and define $A_m = \bigcup_{n\ge m} \{|X_n - X| \ge \epsilon\}$. Then note that for every i, we have $A_i \supseteq X_{i+1}$. Hence $A_\infty = \bigcup_{m\ge 1} A_m$. Then $A_\infty^c = \bigcup_{m\ge 1} A_m^c$ by DeMorgan's law. By the definition of A_m, we obtain $A_\infty^c = \bigcup_{m\ge 1} \bigcap_{n\ge m} \{|X_n - X| \le \epsilon\} = \bigcup_{m\ge 1} \bigcap_{n\ge m} \{\omega\colon \omega \in \Omega,\ |X_n(\omega) - X(\omega)| < \epsilon\}$. Since this quantity is equal to the last equation in the first paragraph, we obtain $\mathbb{P}[A_\infty^c] = 1$, so $\mathbb{P}[A_\infty] = 0$.

Thus, $\mathbb{P}[|X_n - X| \ge \epsilon] \le \mathbb{P}[A_\infty] = 0$, so $\mathbb{P}[|X_n - X| \ge \epsilon] = 0$ for any $n > \max_\omega m(\omega)$. Thus, $\{X_n\} \overset{\text{prob.}}{\to} X$. □

Cleaned Up Proof. Note that, for the definition of almost sure convergence,

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

In particular, when we are computing this probability, we are really summing the probabilities over all ω.

For a sequence $\{X_n\} \overset{\text{a.s.}}{\to} X$, let $S = \{\omega \in \Omega\colon \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\}$. Then by the definition of almost sure convergence above, $\mathbb{P}[S] = 0$.

Fix an $\epsilon > 0$. Let $A_n = \bigcup_{m\geq n}\{|X_m - X| \geq \epsilon\}$. One sees that $A_i \supseteq A_{i+1}$ for all i, so the $\{A_n\}$ are decreasing. One further sees that the sequence decreases to the set $A_\infty = \bigcap_{n\geq 1} A_n$. By the probability axioms, since $A_i \subseteq A_{i+1}$, so $\mathbb{P}[A_i] \geq \mathbb{P}[A_{i+1}]$. Their probabilities thus decrease towards $\mathbb{P}[A_\infty]$. Pick a sample point $\omega \in S^c$; then $\lim_{n\to\infty} X_n(\omega) = X(\omega)$. By the definition of the limit, there exists N such that for all $n > N$, we have $|X_n(\omega) - X(\omega)| < \epsilon$. Therefore, for each $n > N$, $\omega \notin A_n$, so $\omega \notin A_\infty$. Since this works for each point $\omega \in S^c$, we know that $A_\infty \cap S^c = \emptyset$. Equivalently, $A_\infty \subseteq S$. Thus, $\mathbb{P}[A_\infty] = 0$.

Finally, consider $\lim_{\to\infty} \mathbb{P}[|X_n - X| > \epsilon] \leq \lim_{n\to\infty} \mathbb{P}[A_n] = 0$. By definition, this means that $\{X_n\}$ converge in probability to X. □

Example 136 (Convergence in rth moment implies convergence in probability). Let $X_n \sim \text{Bernoulli}(1/n)$. Then $\mathbb{E}[|X_n|^r] = \frac{1}{n}$. In the limit this is 0, so expanding out the limit definition gives that $X_n \overset{\text{prob.}}{\to} X$.

Example 137 (Convergence in probability doesn't imply convergence in rth moment). Let $X_n \sim n^2 \cdot \text{Bernoulli}(1/n)$. Then for any $\epsilon > 0$ we have $\lim_{n\to\infty} \mathbb{P}[|X_n| \geq \epsilon] = \lim_{n\to\infty} \frac{1}{n} = 0$, so $X_n \overset{\text{prob.}}{\to} 0$. But $\mathbb{E}[X_n] = \mathbb{E}[|X_n|] = n$, so the rth moment clearly diverges.

Example 138 (Convergence in rth moment doesn't imply almost sure convergence). Let $X_n \sim \text{Bernoulli}(1/n)$. This converges in the rth moment. We attempt now to compute $\mathbb{P}[\lim_{n\to\infty} |X_n| = 0]$. In particular, for every $\epsilon > 0$ there exists m such that $\mathbb{P}[|X_n - 0| < \epsilon\ \forall\ n \geq m]$. This probability is equal to the quantity $\lim_{n\to\infty} \prod_{i=m}^n \left(1 - \frac{1}{i}\right) = \lim_{n\to\infty} \prod_{i=m}^n \frac{i-1}{i}$. By telescoping, we obtain that $\lim_{n\to\infty} \prod_{i=m}^n \frac{i-1}{i} = \lim_{n\to\infty} \frac{m-1}{n} = 0$. Thus, $\mathbb{P}[\lim_{n\to\infty} |X_n| = 0] = 0$, so X_n does not converge almost surely to X.

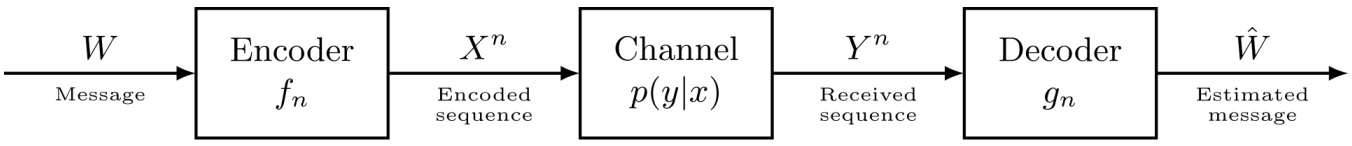
Theorem 139 (Lyapunov's Inequality). If $r > s \geq 1$, then

$$\mathbb{E}[|X_n - X|^s]^{1/s} \leq \mathbb{E}[|X_n - X|^r]^{1/r}$$

From this it's easy to show that convergence in r^{th} moment implies convergence in s^{th} moment.

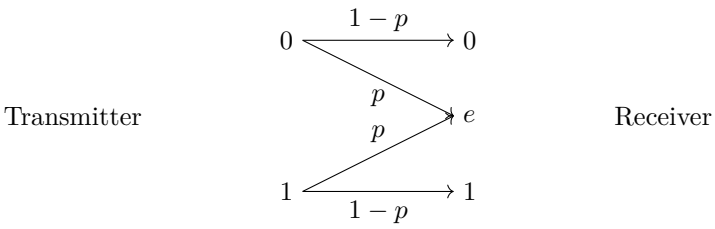
4 Information Theory

The following is Shannon's imagined information theory infrastructure.

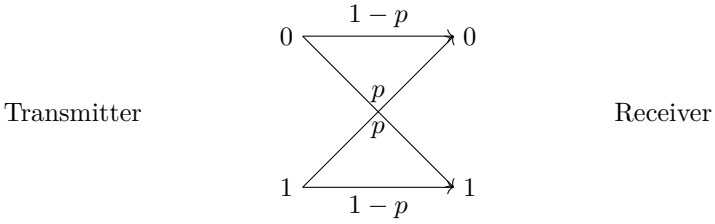


Theorem 140 (Source Coding Theorem). If there are N i.i.d. random variables $X_1, \ldots, X_n$, each with entropy $H(X)$, then they can be compressed into at least $N \cdot H(X)$ bits with negligible risk of information loss, as $N \to \infty$. But, conversely, if they are compressed into fewer than $N \cdot H(X)$ bits, it's almost certain that information will be loss.

There are two main models of sending information over a channel. They are the binary erasure channel:



and the binary symmetric channel:



Definition 141 (Channel Rate and Capacity). The channel rate is defined as

$$c = \frac{\#\text{ of the message input bits}}{\#\text{ of the bits transmitted over the channel}}$$

The capacity of a channel is the highest possible rate.

Definition 142 (Achievable Rate). We say a rate r is achievable for the channel if for each positive integer n there exists a channel encoding and a channel decoding function pair (f_n, g_n), which encode a message of length $L(n) = \lfloor nr \rfloor$ to a message of length n such that $\lim_{n\to\infty} P_e(n) = 0$

Theorem 143 (Channel Coding Theorem). Any rate below the channel capacity c is achievable. Conversely, any sequence of codes with $\lim_{n\to\infty} P_e(n) = 0$ (where $P_e(n)$ is the maximum probability of errors over possible input messages for the channel) has rate $r < c$.

The Channel Coding Theorem has to deal with the behavior of the channel, and its encoder/decoder; the Source Coding Theorem has to deal with the behavior of the source encoder and/or decoder.

Fact 144. The capacity of a binary erasure channel with error probability p is $1 - p$.

Proof. Let $\mathcal{X}$ be the alphabet at channel input; in this case, $\mathcal{X} = \{0, 1\}$. Let $\mathcal{Y}$ be the alphabet at channel output; in this case, $\mathcal{Y} = \{0, 1, e\}$. Let $f_N \colon \mathcal{X}^L \to \mathcal{X}^N$ be the channel encoding function, and $g_N \colon \mathcal{Y}^N \to \mathcal{Y}^L$ be the channel decoding function; these define our encoder and decoder, and thus obey the source coding theorem. Then the maximum probability of error

$$P_e(n) = \max_{x \in \mathcal{X}^L} \mathbb{P}\left[g_N(y^{(k)} \neq x \mid x^{(k)} = f_N(x))\right]$$

Recall that if r is an achievable rate then for every N there exists (f_N, g_N) which encode a message of length $L(N) = \lfloor Nr \rfloor$ to a message of length N such that $\lim_{n\to\infty} P_e(N) = 0$.

If we transmit a bit, we can get an erasure; the best strategy in this case is to immediately retransmit the bit. The number of bits we send to avoid an erasure is distributed according to Geometric$(1 - p)$, so on average we send $\frac{1}{1-p}$ bits. Thus the rate is upper bounded by $r \leq \frac{1}{\frac{1}{1-p}} = 1 - p$.

Now we lower bound r. For an $\epsilon > 0$, define $r = 1 - p - \epsilon$. Consider $2^{L(N)}$ codewords $C_1, \ldots, C_{2^{L(N)}}$, and send N bits over for each, so we send $N2^{L(N)}$ bits in total. For N sufficiently large, we invoke the strong law of large numbers, and work with the means of every variable. The proportion that will appear is $1 - p$, so in each channel we

receive $N(1-p)$ bits. Erase the Np erased bits in the first channel from every channel, and work with the remaining $N(1-p)$ bits. Our goal is to find what the probability is that the code is the same among all $N(1-p)$ bits; then it is impossible to decode the codeword since the channel decoder is not one-to-one.

$$\mathbb{P}[\text{error}] = \mathbb{P}\left[\bigcup_{\substack{i=1\\ i\neq k}}^{2^{L(N)}} (C_1 = C_k)\right] \leq \sum_{i=1}^{2^{L(N)}} 2^{-\lfloor N(1-p)\rfloor} = 2^{L(N)} \cdot 2^{-\lfloor N(1-p)\rfloor} \approx 2^{-N\epsilon}$$

and for large N this approaches 0, so any $r \leq 1-p$ is achievable. □

Fact 145. The capacity of a binary symmetric channel with error probability p is $1 - H(p)$.

Proof. Walrand. □

We now discuss source encoding, since most of the previous work is on channel encoding. One scheme for source encoding is Huffman Coding. The idea behind Huffman Coding is to ensure that no code is a prefix of another, so each encoded message is uniquely interpretable, and more likely symbols have smaller bit representations.

The naive approach to coding an alphabet Σ takes $\lg(|\Sigma|)$ bits. We aim to do better than this. Walrand's book goes over Huffman tree construction.

5 Markov and Poisson Processes

Markov Chains

Definition 146. A model has the **Markov Property** if the effect of past states of the system is summarized in the current state.

To prove that a model has the Markov property, you take some arbitrary past state and show how that has unique representation in the current state.

Example 147. Let F_n be a gambler's winnings after n epochs. Let G_n be the amount won if the gambler wins his n^{th} game; let C_n be the value of the n^{th} game. Then

$$F_{n+1} = F_n + G_{n+1} \cdot \Big(\mathbb{1}(C_n = H) - \mathbb{1}(C_n = T)\Big)$$

for $F_n > 0$ (since you can't gamble with no money) and $F_n < L$ (there's a threshold beyond which the gambler goes home and celebrates). If $F_n \leq 0$ or $F_n \geq L$ then $F_{n+1} = F_n$.

Example 148. Let T_n be the time for the n^{th} earthquake to occur in a given area. Let $\{Y_n\} \overset{\text{i.i.d.}}{\sim} \text{Exponential}(1/\lambda)$. Then we can model T_n as

$$T_{n+1} = T_n + Y_n$$

If $X(t)$ is the number of earthquakes by time t, then $X(t)$ is a Poisson process.

A special and important case of Markov processes are Birth and Death processes, where the state can only go forward or backwards by a single amount. Birth and death processes can go up by a or down by b, and a and b can even be time-dependent.

Discrete-Time Markov Chain

Definition 149. A discrete time Markov chain $(X_n)_{n\in\mathbb{N}}$ is a Markov chain that undergoes state changes at discrete times, there are only a finite set of states $S = \{1, \ldots, m\}$, and it satisfies the Markov property for transitions from state $i \in S$ to state $j \in S$:

$$\mathbb{P}[X_{n+1} = j \mid X_n = i, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1] = \mathbb{P}[X_{n+1} = j \mid X_n = i]$$

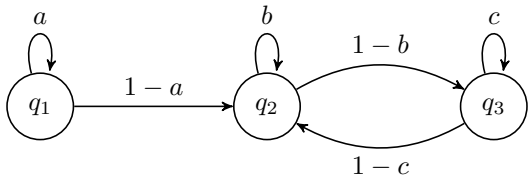
$$= p_{ij}$$

where $p_{ij} \geq 0$, and $\sum_j p_{ij} = 1$. Furthermore, we deal with **time homogeneous** Markov chains: the state change probabilities are not time-dependent.

Definition 150. The probability transition matrix P contains all the information about transitions between different states: $[P]_{ij} = p_{ij}$. The sum over the rows is 1.

Let $\pi^{(n)} = [\mathbb{P}[X_n = 1], \ldots, \mathbb{P}[X_n = m]]$. Then by the Law of Total Probability, $\pi^{(n+1)}(i) = \sum_j \pi^{(n)}(j) p_{ji}$, so $\pi^{(n+1)} = \pi^{(n)} P$, so by induction $\pi^{(n)} = \pi^{(0)} P^n$.

Example 151. Suppose we have the following Markov Chain:



Let's look at how Markov property shows up. We know that

$$\mathbb{P}[X_1, \ldots, X_n] = \prod_{i=1}^{n} \mathbb{P}\left[X_i \mid \bigcap_{j=1}^{i-1} X_j\right]$$

Now by Markov Property,

$$\prod_{i=1}^{n} \mathbb{P}\left[X_i \mid \bigcap_{j=1}^{i-1} X_j\right] = \prod_{i=1}^{n} \mathbb{P}[X_i \mid X_{i-1}]$$

which is how we compute the joint probabilities.

Let $r_{ij}(n) = \mathbb{P}[X_n = j \mid X_0 = i]$ represent the probability that we are in state j exactly n steps after reaching state i. The value of $r_{ij}(n)$ can be calculated recursively as

$$r_{ij}(n) = \sum_{k \in S} r_{ik}(n-1) p_{kj}$$

We can show by induction that $r_{ij}(n) = [P^n]_{ij}$.

Definition 152 (Accessibility and Recurrence). A state j is **accessible** or **reachable** from state i if there exists $n \in \mathbb{N}$ such that $r_{ij}(n) > 0$.

A state i is **recurrent** if for all j that are reachable from i, i is reachable from j. Formally, if $A(i)$ is the set of reachable states from i, then i is recurrent if for all $j \in A(i)$, $i \in A(j)$.

A state i is **transient** if it is not recurrent.

Example 153. For the Markov Chain example above, we assume $b, c \neq 0$. If $a = 1$, then q_1, q_2, q_3 are recurrent. If $a < 1$, then q_1 is transient and q_2, q_3 are recurrent.

Definition 154 (Recurrence Class). For any recurrent state i, all states $A(i)$ (the set of states reachable from i) form a **recurrent class**. Any Markov chain can be decomposed into one or more recurrent classes.

These are analogous to strongly connected components in directed graphs. Analogously, a state in a recurrent class is not reachable from classes in any other recurrence class.

Definition 155 (Irreducible). A Markov chain is called **irreducible** if it only has one recurrent class. For any non-irreducible Markov chain, we can identify the recurrent classes by finding the strongly connected components of the graph depicted by the adjacency matrix P.

Definition 156 (Periodicity). Consider an irreducible Markov chain. For a recurrent state $i \in S$, define

$$d(i) = \gcd(()\{n \geq 1 \colon r_{ii}(n) > 0\})$$

In other words, all paths from i back to i take a multiple of $d(i)$ steps.

Fact 157. For all states i in the same recurrent class, $d(i)$ is the same. For Markov chains with more than one recurrent class, each class has a separate value for d, as each class is a 'sub-Markov chain'. We define an irreducible Markov chain as **aperiodic** if for every recurrent state i, $d(i) = 1$. Otherwise, we say that it's periodic with period $d(i)$ for any recurrent state i.

Note that we don't really define $d(i)$ for transient states $d(i)$.

Corollary 158. Adding a self-loop will make an irreducible Markov chain aperiodic.

Definition 159 (Stationary Distribution). If we choose the initial state of the Markov chain according to the distribution

$$\mathbb{P}[X_0 = j] = \pi^{(0)}(j) \qquad \forall\, j \in S$$

and this implies

$$\mathbb{P}[X_n = j] = \pi^{(0)}(j) \qquad \forall\, j \in S, n \in \mathbb{N}$$

then we say that $\pi^{(0)}$ is **stationary**.

In other words, if π is some distribution on X_i for some i, then if for all $n \geq 0$ we have that $\pi P^n = \pi$, then π is stationary. Accordingly, if we verify that for all j we have

$$\pi(j) = \sum_{k=1}^{m} \pi(k) p_{kj}$$

then π is a stationary distribution. The final equivalent condition is that π is a left eigenvector of P that has corresponding eigenvalue $\lambda_\pi = 1$. In general, there can be multiple unique stationary distributions; if we have two stationary distributions, then we can get infinitely many stationary distributions due to taking convex combinations; the number of linearly independent stationary distributions is bounded above by $\operatorname{rank}(()P)$.

For irreducible Markov chains, there must exist a stationary distribution.

By subtracting $\pi(i)p_{ii}$ from both sides of the balance equation, we have for all states i in a stationary distribution

$$\underbrace{\sum_{j \neq i} \pi(j) p_{ji}}_{\text{probability flow in}} = \underbrace{\pi(i) \sum_{j \neq i} p_{ij}}_{\text{probability flow out}}$$

In particular, if we take a vertex cut of the graph associated with P, then the flow inside the cut must equal the flow outside of the cut. This is a good interpretation of how a Markov chain is just the visualization of a probability flow between states.

Theorem 160 (Hitting Times on First Passage Times). Let $A \subseteq S$. Let $T_A = \min(\{n \geq 0 \mid X_n \in A\})$. Then

$$\mathbb{E}[T_A \mid X_0 = i] = \left(1 + \sum_{j \in S} p_{ij} \mathbb{E}[T_A \mid X_0 = j]\right) \cdot \mathbb{1}(i \notin A)$$

Example 161. Given a discrete time Markov chain with $S = \{0, 1, 2, 3\}$ and probability transition matrix

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1-p & 0 & p & 0 \\ 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

with $A = \{0, 3\}$, then $\beta(i) = \mathbb{E}[T_A \mid X_0 = i]$ has $\beta(0) = 0$, $\beta(3) = 0$, $\beta(1) = 1 + p\beta(2)$, and $\beta(2) = 1 + (1-p)\beta(1)$, so $\beta(1) = (1+p)/(1-p+p^2)$ and $\beta(2) = (2-p)/(1-p+p^2)$.

Theorem 162. Let $A \subseteq S$ and $B \subseteq S$ with $A \cap B = \emptyset$. Let T_A and T_B be defined according to the previous theorem. Then

$$\mathbb{P}[T_A < T_B \mid X_0 = i] = \left(\sum_{j \in S} p_{ij} \mathbb{E}[T_A < T_B \mid X_0 = j] \right) \cdot \mathbb{1}(i \notin A \cup B) + \mathbb{1}(i \in A)$$

Example 163. Given a discrete time Markov chain with $S = \{0, 1, 2, 3\}$ and probability transition matrix

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1-p & 0 & p & 0 \\ 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

with $A = \{0\}$ and $B = \{3\}$, then $\alpha(i) = \mathbb{P}[T_A > T_B \mid X_0 = i]$ has $\alpha(0) = 0$, $\alpha(3) = 1$, $\alpha(1) = p\alpha(2)$ and $\alpha(2) = p + (1-p)\alpha(1)$, so $\alpha(1) = p^2/(1-p+p^2)$ and $\alpha(2) = p/(1-p+p^2)$.

Before, we have been dealing with cases where the state space S is finite, i.e. $|S| < |\mathbb{N}|$. Now we deal with the case where the state space S is countably infinite, i.e. $|S| = |\mathbb{N}|$.

Definition 164 (Recurrence and Transience for Discrete Time Markov Chains with Countably Infinite States). Define $T_x = \min(\{n \geq 0 \colon X_n = x\})$ and $T_x^+ = \min(\{n \geq 1 \colon X_n = x\})$. For $x, y \in S$, let $\rho_{x,y} = \mathbb{P}\left[T_y^+ < \infty \mid X_0 = x\right]$, and let $\rho_x = \rho_{x,x}$. We say x is **recurrent** if $\rho_x = 1$ and **transient** if $\rho_x < 1$.

Corollary 165. Let $N_k(x) = \sum_{n=0}^{k} \mathbb{1}(X_n = x)$.

If x is recurrent then the sequence $(N_n(x))_{n \in \mathbb{N}} \overset{\text{a.s.}}{\to} \infty$ under the conditional probability law $\mathbb{P}[\cdot \mid X_0 = x]$, that is,

$$\mathbb{E}\left[\lim_{n \to \infty} N_n(x) \mid X_0 = x\right] = \infty$$

that is, x is visited an infinite number of times in the progression of the Markov chain.

If x is transient then the sequence $(N_n(x))_{n \in \mathbb{N}} \overset{\text{a.s.}}{\to} \rho_x/(1-\rho_x)$ under the conditional probability law $\mathbb{P}[\cdot \mid X_0 = x]$, that is,

$$\mathbb{E}\left[\lim_{i \to \infty} N_i(x) \mid X_0 = x\right] = \frac{\rho_x}{1-\rho_x}$$

that is, x is visited a finite number of times in the progression of the Markov chain.

Proof. The first part is trivial.

For the second part, by the tail-sum formula,

$$\mathbb{E}\left[\lim_{n \to \infty} N_n(x) \mid X_0 = x\right] = \sum_{k \in \mathbb{N}} \mathbb{P}\left[\lim_{n \to \infty} N_n(x) \geq k \mid X_0 = x\right] = \sum_{k \in \mathbb{N}} \rho_x^{k+1} = \frac{\rho_x}{1-\rho_x}$$

as desired. □

We see that these definitions are stronger than those for the finite-state discrete time Markov chain, that is, the finite-state definitions do not imply the infinite-state definitions, but the infinite-state definitions imply the finite-state definitions.

Example 166 (Random Walk). Let $(X_n)_{n \in \mathbb{N}}$ be a Markov chain with state space $S = \mathbb{N}$. Furthermore, let $\mathbb{P}[X_i = 0 \mid X_{i-1} = 0] = \mathbb{P}[X_i = j \mid X_{i-1} = j-1] = \mathbb{P}[X_i = j \mid X_{i-1} = j+1] = 1/2$ for $j > 0$. Then we attempt to find ρ_0. Immediately from the recurrence we obtain $\rho_0 = \frac{1}{2} + \frac{1}{2}\rho_{1,0}$ and $\rho_{1,0} = \frac{1}{2} + \frac{1}{2}\rho_{2,0}$. By the Markov property, we have $\rho_{2,0} = \rho_{2,1} \cdot \rho_{1,0} = \rho_{1,0}^2$ (in general, $\rho_{i,i-1} = \rho_{1,0}$). We obtain $\rho_{1,0} = \frac{1}{2} + \frac{1}{2}\rho_{1,0}^2$, so $\rho_{1,0} = 1$, so $\rho_0 = 1$. Thus, for this random walk, 0 is recurrent, and by the previous corollary, $(N_n(0))_{n \in \mathbb{N}} \overset{\text{a.s.}}{\to} \infty$.

Proposition 167. A finite-state discrete time Markov chain must have at least one recurrent state.

Proof. Since there are finite states but infinite state-changes, by Pidgeonhole principle at least one state must be returned to infinitely many times with probability 1. □

Definition 168. A state x **communicates** with state y if $\rho_{x,y} > 0$ and $\rho_{y,x} > 0$ (which is the same as $y \in A(x)$ and $x \in A(y)$). A **communicating class** is a maximal set of states that communicate with each other.

Definition 169. A discrete time Markov chain is irreducible if it consists of only a single communicating class.

Communicating classes are weaker than recurrence classes; a recurrence class is a communicating class.

Theorem 170. All states in a given communicating class are either recurrent or transient.

Definition 171. State x is **positive recurrent** if it's recurrent and $\mathbb{E}[T_x^+ \mid X_0 = x] < \infty$. State x is **null recurrent** if it's recurrent and $\mathbb{E}[T_x^+ \mid X_0 = x] = \infty$.

Theorem 172 (Ergodic Theorem for Markov Chains). If $(X_n)_{n\in\mathbb{N}}$ is a discrete time irreducible Markov chain with state space S, where $|S| \leq |\mathbb{N}|$, then

- if the Markov chain is positive recurrent, then there exists a unique stationary distribution.
- if there exists a stationary distribution, then the Markov chain is positive recurrent (and the stationary distribution is unique).
- if the Markov chain is irreducible and positive recurrent, then $\frac{1}{n}\sum_{i=0}^{n-1} \mathbb{1}(X_i = x) \overset{\text{a.s.}}{\to} \pi(x)$, that is
$$\mathbb{P}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{1}(X_i = x) = \pi(x)\right] = 1.$$
- if the Markov chain is irreducible, positive recurrent, and aperiodic, then $\lim_{n\to\infty} \mathbb{P}[X_n = x] = \pi(x)$.

Note that an irreducible finite-state Markov chain is positive recurrent.

Theorem 173. Consider an irreducible and positive recurrent discrete-time Markov chain with stationary distribution π. Then
$$\pi(x) = \frac{1}{\mathbb{E}\left[T_x^+ \mid X_0 = x\right]}$$

Proof Sketch. Let $\tau_1, \tau_2, \ldots$ be the inter-visit intervals for x. Then by the Markov property, $\tau_1, \tau_2, \ldots$ are i.i.d.. Then by the Strong Law of Large Numbers,
$$\frac{1}{n}\sum_{i=1}^{n} \tau_i \overset{\text{a.s.}}{\to} \mathbb{E}\left[T_x^+ \mid X_0 = x\right]$$
Then take some large time $t \geq 0$; we have
$$\lim_{t\to\infty} \frac{t}{\sum_{i=0}^{t-1} \mathbb{1}(X_i = x)} = \mathbb{E}\left[T_x^+ \mid X_0 = x\right]$$
Inverting obtains, using Dominated Convergence Theorem,
$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=0}^{t-1} \mathbb{1}(X_t = x) \overset{\text{a.s.}}{\to} \frac{1}{\mathbb{E}\left[T_x^+ \mid X_0 = x\right]}$$
But by the Ergodic Theorem we know that the left-hand quantity almost surely converges to $\pi(x)$, so
$$\pi(x) = \frac{1}{\mathbb{E}\left[T_x^+ \mid X_0 = x\right]}$$
as desired. □

Corollary 174 (General Random Walk). Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with state space $S = \mathbb{N}$. For some $p \in [0,1]$, let $\mathbb{P}[X_i = 0 \mid X_{i-1} = 0] = \mathbb{P}[X_i = j \mid X_{i-1} = j+1] = 1 - p$, and let $\mathbb{P}[X_i = j \mid X_{i-1} = j-1] = p$ for

$j > 0$. Then if $p > \frac{1}{2}$, all states are transient; if $p < \frac{1}{2}$, all states are positive recurrent; if $p = \frac{1}{2}$, all states are null recurrent.

Proof. We show previously in the notes that if $p = \frac{1}{2}$ then the Markov chain is recurrent; we now specifically show that it's recurrent. The first-step equations give

$$\mathbb{E}\left[T_0^+ \mid X_0 = 1\right] = 1 + \frac{1}{2}\mathbb{E}\left[T_0^+ \mid X_0 = 2\right] = 1 + \frac{1}{2}\left(\mathbb{E}\left[T_1^+ \mid X_0 = 2\right] + \mathbb{E}\left[T_0^+ \mid X_0 = 1\right]\right)$$

By symmetry, if $p = \frac{1}{2}$, then $\mathbb{E}\left[T_1^+ \mid X_0 = 2\right] = \mathbb{E}\left[T_0^+ \mid X_0 = 1\right]$, so $\mathbb{E}\left[T_0^+ \mid X_0 = 1\right] = 1 + \mathbb{E}\left[T_0^+ \mid X_0 = 1\right]$, so $0 = 1$, contradiction, so $\mathbb{E}\left[T_0^+ \mid X_0 = 1\right] = \infty$.

Now we show that if $p > \frac{1}{2}$, then the Markov chain is transient. Let $Y_n \overset{\text{i.i.d.}}{\sim} 2 \cdot \text{Bernoulli}(p) - 1$. Then $X_n = \max\left(X_{n-1} + Y_n, 0\right)$, for $n \geq 1$. Then $\mathbb{P}[\lim_{n\to\infty} X_n \geq \lim_{n\to\infty} X_0 + \sum_{i=1}^n Y_i]$, so

$$\mathbb{P}\left[\lim_{n\to\infty} \frac{1}{n} X_n \geq \lim_{n\to\infty} \left(\frac{X_i}{n} + \frac{1}{n}\sum_{i=1}^n Y_i\right)\right]$$

so $\mathbb{P}[X_n = 0] \overset{\text{a.s.}}{\to} 0$. Hence 0 is transient, so the Markov chain is transient. □

Before we have been dealing with global balance equations; now we introduce local balance equations.

Definition 175 (Local Balance Equations). Let S be a cut of the Markov graph. Then

$$\sum_{(i,j)\colon i\in S, j\in V\setminus S} \pi(i)p_{ij} = \sum_{(j,i)\colon j\in V\setminus S, i\in S} \pi(j)p_{ji}$$

i.e. the flow into a cut equals the flow out of a cut.

Here's an example where this approach makes sense.

Example 176 (Birth-Death Markov Chain). Take a discrete-time Markov chain where $\mathbb{P}[X_n = 0 \mid X_{n-1} = 0] = d_0$, $\mathbb{P}[X_{n+1} = i \mid X_n = i-1] = b_i$, and $\mathbb{P}[X_{n+1} = i-1 \mid X_n = i] = d_i$. Take a cut of states $\{0, \ldots, k\}$. Then by this interpretation, $\pi(k)b_k = \pi(k+1)d_{k+1}$, so $\pi(k) = \pi(0)\prod_{i=0}^{k-1} \frac{b_i}{d_{i+1}}$.

For a finite random walk, if $b_i = p$ and $d_i = 1 - p$, then for $p < \frac{1}{2}$, then $\pi(k) = \pi(0)\left(\frac{p}{1-\square}\right)^l$, so $\pi(k) = \left(1 - \frac{p}{1-p}\right)\left(\frac{p}{1-p}\right)^k$.

Definition 177 (Detailed Balance Equations). For any states i and j, $\pi(i)p_{ij} = \pi(j)p_{ji}$.

These imply the regular balance equations.

Definition 178 (Reversibility). Let $(X_n)_{n\in\mathbb{N}}$ be an irreducible, positive recurrent discrete-time Markov chain with stationary distribution π. Let $\pi_0 = \pi$. If for all $n \geq 0$, $p_{X_0,X_1,\ldots,X_n}(\cdot) = p_{X_n,X_{n-1},\ldots,X_0}(\cdot)$, then $(X_n)_{n\in\mathbb{N}}$ is reversible.

Fact 179. The reverse discrete-time Markov chain has the Markov property.

Proof. By computing the conditional probability:

$$\begin{aligned}
\mathbb{P}[X_k = i \mid X_{k+1} = j, X_{k+2} = i_{k+2}, \ldots, X_n = i_n] &= \frac{\mathbb{P}[X_k = i, X_{k+1} = j, X_{k+2} = i_{k+2}, \ldots, X_n = i_n]}{\mathbb{P}[X_{k+1} = j, X_{k+2} = i_{k+2}, \ldots, X_n = i_n]} \\
&= \frac{\pi(i)p_{ij}p_{ji_{k+2}}\prod_{\ell=k+2}^{n-1} p_{i_\ell i_{\ell+1}}}{\pi(j)p_{ji_{k+2}}\prod_{\ell=k+2}^{n-1} p_{i_\ell i_{\ell+1}}} \\
&= \frac{\pi(i)p_{ij}}{\pi(j)} \\
&= p_{ji} \quad \text{(if reversible)}
\end{aligned}$$

This just depends on the candidate state i and the previous state j, so this has the Markov property. □

In particular, for reversibility $\mathbb{P}[X_{k+1} = i \mid X_k = j] = p_{ji}$. So $\pi(i)p_{ij} = \pi(j)p_{ji}$ implies reversibility. So if the detailed balance equations hold the Markov chain is reversible. This means that the birth-and-death chain is reversible.

Theorem 180. If π satisfies the detailed balance equations, then π is a stationary distribution.

The reverse Markov chain has the same stationary distribution π as the normal Markov chain.

Example 181. For the Birth-Death Markov chain, if $b_i = p$ and $d_i = 1-p$, then $\mathbb{P}[X_0 = 2, X_1 = 3, X_2 = 2, X_3 = 1] = \left(1 - \frac{p}{1-p}\right)\left(\frac{p}{1-p}\right)^2 \cdot p \cdot (1-p) \cdot (1-p)$. The reverse Markov chain has this same probability; if Y is the reverse Markov chain, then $\mathbb{P}[Y_3 = 1, Y_2 = 2, Y_3 = 3, Y_4 = 2] = \left(1 - \frac{p}{1-p}\right) \cdot \frac{p}{1-p} \cdot p \cdot p \cdot p^2$; these are actually the same quantity.

Poisson Processes

Definition 182 (Poisson Process). An arrival process $\{N(t), t \geq 0\}$ is called a **Poisson process** with rate $\lambda > 0$ if it obeys the properties:

- Time-homogeneity: $\text{Poisson}_\lambda(k, \tau) = \mathbb{P}[k \text{ arrivals in } I]$ is the same for any interval I of length τ.
- Independence: the number of arrivals in a given interval is independent of what happens outside of the interval.
- $\text{Poisson}_\lambda(0, \tau) = 1 - \lambda\tau + o(\tau)$, where $\lim_{\tau \to 0} \frac{o(\tau)}{\tau} = 0$. Similarly, $\text{Poisson}_\lambda(1, \tau) = \lambda\tau + o_1(\tau)$, where $o_1(\tau)$ has the same limiting behavior as $o(\tau)$. In general, $\text{Poisson}_\lambda(k, \tau) = o_k(\tau)$ for $k > 1$.
- We have $\text{Poisson}_\lambda(k, \tau) = \frac{\mathrm{e}^{-\lambda\tau}(\lambda\tau)^k}{k!}$.

Recall the Poisson approximation to the binomial: if $Z \sim \text{Poisson}(\lambda)$ and $S \sim \text{Binomial}(n, p)$, then if $\lim_{n\to\infty} np = \lambda$, we have $\lim_{n\to\infty} p_S(k) = p_Z(k)$.

In the same way, imagine that we have a time interval $I(\tau) = [0, \tau]$. Divide $I(\tau)$ into subintervals $\phi(i)$ given by $[0, \frac{\tau}{n}) \cup [\frac{\tau}{n}, 2\frac{\tau}{n}) \cup \cdots \cup [(n-1)\frac{\tau}{n}, \tau]$, where $\phi(i) = [i\frac{\tau}{n}, (i+1)\frac{\tau}{n})$. Let X_i be the number of arrivals in the interval $\phi(i)$. Then because of our limiting behavior, $\lim_{n\to\infty} \mathbb{P}[X_i = 0] = 1 - \lambda\frac{\tau}{n}$,$\lim_{n\to\infty} \mathbb{P}[X_i = 1] = \lambda\frac{\tau}{n}$, and for each $k > 1$ we have $\lim_{n\to\infty} \mathbb{P}[X_i = k] = 0$. Thus the number of arrivals in $\lim_{n\to\infty} \phi(i)$ is distributed according to $\lim_{n\to\infty} \text{Bernoulli}\left(\lambda\frac{\tau}{n}\right)$.

We now look for the time of the first arrival, starting at $t = 0$. Let T be the time of the first arrival. Then for some time t, we have

$$\begin{aligned}\mathbb{P}[T \leq t] &= 1 - \mathbb{P}[T > t] \\ &= 1 - \underset{\lambda}{\text{Poisson}}(\lambda, t) \\ &= 1 - \mathrm{e}^{-\lambda t}\end{aligned}$$

Thus $T \sim \text{Exponential}(\lambda)$, so $\mathbb{E}[T] = \frac{1}{\lambda}$ and $\text{Var}[T] = \frac{1}{\lambda^2}$.

For arrivals in disjoint intervals, let N_1 be the number of arrivals in τ_1 and N_2 be the number of arrivals in τ_2. Then $N_1 + N_2 \sim \text{Poisson}(\ell(\tau_1) + \ell(\tau_2))$ where $\ell(I)$ is the Lebesgue measure of interval I. This is just Poisson merging and splitting.

There are two useful properties of the Poisson process, starting at $t = t'$.

- For any $t > t'$, the Poisson process after t is independent of the Poisson process before t. (This implies that inter-arrival times are distributed identically and independently).
- For any $t > t'$, let T be the time of the first arrival. Then for some $s > 0$,

  $$\mathbb{P}[T - t > s] = \mathbb{P}[0 \text{ arrivals in } [t, t+s]] = \underset{\lambda}{\text{Poisson}}(0, s) = \mathrm{e}^{-\lambda s}$$

  It's possible to show this using the definition of X_s.

For $i > 0$, let T_i be the inter-arrival time between the $(i-1)^{\text{th}}$ and i^{th} arrivals (T_1 is just the time of the first arrival). We know that $T_i \overset{\text{i.i.d.}}{\sim} \text{Exponential}(\lambda)$. This motivates an alternate definition of the Poisson process for $t \geq t'$: define $T_i \overset{\text{i.i.d.}}{\sim} \text{Exponential}(\lambda)$. Then let T_i be the i^{th} inter-arrival time. This recovers the original Poisson process.

Let Y_k be the k^{th} arrival time. Then $Y_k = \sum_{i=1}^{k} T_i$. Then $\mathbb{E}[Y_k] = k\mathbb{E}[T_i] = \frac{k}{\lambda}$. Similarly, $\text{Var}[Y_k] = \frac{k}{\lambda^2}$.

Fact 183. The actual distribution of Y_k is given by

$$p_{Y_k}(y) = \frac{\lambda^k y^{k-1} \mathrm{e}^{-\lambda y}}{(k-1)!}$$

which is a distribution called the Γ-distribution, or the Erlang distribution of order k.

Proof. Computing the cumulative density function,:

$$P_{Y_k}(y) = \mathbb{P}[Y_k \leq y] = \sum_{n=k}^{\infty} \underset{\lambda}{\text{Poisson}}(n, y) = 1 - \sum_{n=0}^{k-1} \underset{\lambda}{\text{Poisson}}(n, y)$$

and $f_{Y_k}(y) = \frac{\mathrm{d}}{\mathrm{d}y} P_{Y_k}(y)$. □

Now we consider splitting and merging.

Fact 184 (Splitting). If P is a Poisson process with rate λ, and for every arrival occurring in P it goes into any one of $n > 0$ subprocesses P_i with probability ϕ_i (if $\sum_{i=1}^{n} \phi_i = 1$), then P_i is a Poisson process with rate $\lambda\phi_i$.

Proof. For all i, the probability that P_i has an arrival in a given interval of length $\frac{\tau}{n}$ is $\phi_i \lambda \frac{\tau}{n}$. Then the probability of arriving in any of the n intervals is $\phi_i \lambda_i \tau$, so it's a Poisson process with parameter $\phi_i \lambda$. □

Fact 185 (Merging). If P_i are Poisson processes with rate λ_i, and P is a Poisson process that has an arrival when any P_i has an arrival, then P is a Poisson process with rate $\sum_{i=1}^{n} \lambda_i$.

Proof. The probability of 0 arrivals in some interval of length $\frac{\tau}{n}$ is $\prod_{i=1}^{n} \left(1 - \lambda_i \frac{\tau}{n}\right) \approx 1 - \frac{\tau}{n} \sum_{i=1}^{n}$. The probability of 1 arrival in this interval is $\sum_{i=1}^{n} \lambda_i \frac{\tau}{n} \prod_{j \neq i} (1 - \lambda_j \cdot \frac{\tau}{n}) \approx \frac{\tau}{n} \sum_{i=1}^{n} \lambda_i$. Thus the merged process is a Poisson process with parameter $\sum_{i=1}^{n} \lambda_i$. □

Fact 186. If A and B are Poisson processes which merge to P, and T_A is an inter-arrival time from A and T_B is an inter-arrival time from B, then T_A and T_B are independent.

Fact 187. If $X_i \sim \text{Exponential}(\lambda_i)$, then $\mathbb{P}[X_i = \min(X_1, \ldots, X_n)] = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}$

Proof. A homework problem. □

Example 188 (Residual Life Paradox). Let P be a Poisson process starting at time $-\infty$ with parameter λ. Let $t^* \in \mathbb{R}$, and let u be the time of the last arrival, and v be the time of the next arrival. Let $L = (t^* - u) + (v - t^*)$. Due to the independence property of inter-arrival times, $t^* - u$ is independent of $v - t^*$. Memorylessness implies that $v - t^* \sim \text{Exponential}(\lambda)$. Thus

$$\mathbb{P}[t^* - u > x] = \mathbb{P}[\text{no arrival in } [t^* - x, t^*)] = \underset{\lambda}{\text{Poisson}}(0, x) = \mathrm{e}^{-\lambda x}$$

Similarly $t^* - u \sim \text{Exponential}(\lambda)$, so L is an Erlang distribution of order 2. Thus $\mathbb{E}[L] = \frac{2}{\lambda}$. This is because, if we pick a random time, it's more likely to be in a longer inter-arrival time. In general, if the inter-arrival times X_i are i.i.d., then $\mathbb{E}[L] = \frac{\mathbb{E}[X_i^2]}{\mathbb{E}[X_i]}$.

If $\{N(t), t \geq 0\}$ is a Poisson process with rate λ, and given $\{N(s), s \leq t\}$, then $\{N(s+t) - N(t), s \geq 0\}$ is a Poisson process with parameter λ.

One fact that is useful for computing $\mathbb{E}[Y \mid X]$ is if $Y = \sum_{i=1}^{n} X_i$ and X_i are i.i.d., then $\mathbb{E}[X_i \mid S] = \frac{S}{n}$. Furthermore, $X_i \mid S$ is uniform, although the parameter may be nontrivial to compute.

Continuous-Time Markov Chain

Let $\mathcal{X}$ be a finite or countably infinite state space, and let π be a probability distribution on $\mathcal{X}$. Define a function Q by the properties $Q = \{Q(i,j) \colon i, j \in \mathcal{X}\}$ such that $Q(i,j) \geq 0$ for each i and $j \neq i$, and furthermore $\sum_j Q(i,j) = 0$ for all i. Thus $\sum_{j \neq i} Q(i,j) = -Q(i,i)$. Finally, define $q(i) = -Q(i,i)$ for each i. A continuous-time Markov chain over $\mathcal{X}$ with initial distribution π and rate function Q is the process $\{X_t, t \geq 0\}$ such that

$$\mathbb{P}[X_0 = i] = \pi(i)$$

and for each $\epsilon > 0$ we have

$$\mathbb{P}[X_{t+\epsilon} = j \mid X_t = i, X_u \ \forall \ u < t] = \mathbb{P}[X_{t+\epsilon} = j \mid X_t = i] = (1 - q(i)\epsilon) \cdot \mathbb{1}(j = i) + Q(i,j)\epsilon \cdot \mathbb{1}(j \neq i) + o(\epsilon)$$

Note that

$$\sum_j \mathbb{P}[X_{t+\epsilon} = j \mid X_t = i, X_u \ \forall \ u < t] \approx 1$$

and indeed becomes 1 in the limit, which is obvious. An intuitive description for Q is that $Q(i,j)$ is the probability of jumping from i to j in unit time.

We now construct the Markov chain. Suppose $X_t = i$. Let $\tau \sim \text{Exponential}(q(i))$. At time $t + \tau$, the process jumps to $j \neq i$ with probability $\Gamma(i,j) = \frac{Q(i,j)}{q(i)}$ for all $j \neq i$.

For $j \neq i$ and $\epsilon > 0$ small, observe

$$\mathbb{P}[X_{t+\epsilon} = j \mid X_t = i] = \underbrace{q(i)\epsilon}_{\mathbb{P}[X_{t+\epsilon} \neq i \mid X_t = i]} \cdot \underbrace{\frac{Q(i,j)}{q(i)}}_{\mathbb{P}[X_{t+\epsilon} = j \mid X_{t+\epsilon} \neq i, X_t = i]} = \epsilon \cdot Q(i,j)$$

One can justify this choice for $\tau \sim \text{Exponential}(q(i))$ by alternatively setting up a Poisson process $\{N_j(s), s \geq t\}$ for each choice of state that state i can transition to with rate $Q(i,j)$, then the actual transition is just the first arrival in any of the Poisson processes; this is equivalent to finding $\min(\{T_j\})$, which is distributed according to $\text{Exponential}\left(\sum_j Q(i,j)\right)$.

The discrete-time Markov chain $\{X_n, n \geq 0\}$ with transition matrix Γ is called the embedded Markov chain.

Example 189. The Poisson process $\{N_\lambda(t), t \geq 0\}$ with $\pi(0) = 1$ has rate matrix $Q(i,i) = -\lambda$ and $Q(i, i+1) = \lambda$. The transition matrix Γ just has $\Gamma_{i,j} = \mathbb{1}(j = i+1)$.

Example 190 (Machine Repairs). Consider two machines, where the time from when they start working to the time they fail is distributed i.i.d. as $\text{Exponential}(\lambda)$, and from a failing state to a working state the time is distributed i.i.d. as $\text{Exponential}(\mu)$. Let (i,j) be the state where $i = \mathbb{1}(\text{Machine 1 working})$ and $j = \mathbb{1}(\text{Machine 2 working})$. Let $s(i,j) = i + j$ be the states, then $Q(0,1) = Q(1,2) = \mu$ and $Q(2,1) = Q(1,0) = \lambda$. On the other hand, the Γ matrix is a bit more complex, but easily derivable.

Example 191 (Queue). Consider a queue where people arrive in the queue according to a Poisson process with parameter λ (so the time for a person to arrive is distributed according to a random variable distributed $\text{Exponential}(\lambda)$), and the service times are distributed according to a random variable distributed $\text{Exponential}(\mu)$. Then the Markov chain has $Q(0,1) = \lambda$, $Q(0,0) = -\lambda$, and for all $i > 0$ we have $Q(i, i-1) = \mu$, $Q(i,i) = -\lambda + \mu$, and $Q(i, i+1) = \lambda$. The embedded discrete-time Markov chain has $P(0,1) = 1$, and for each $i > 0$ we have $P(i, i-1) = \frac{\mu}{\lambda + \mu}$ and $P(i, i+1) = \frac{\lambda}{\lambda + \mu}$ (think of the basketball problem from homework).

We will try to expand our definitions from the discrete-time Markov chain to the continuous-time Markov chain.

Definition 192. The continuous-time Markov chain is irreducible if and only if the embedded discrete-time Markov chain is irreducible.

Definition 193. A state is recurrent in the continuous-time Markov chain if and only if it's recurrent in the embedded discrete-time Markov chain. Similarly, a state is transient in the continuous-time Markov chain if and only if it's transient in the embedded discrete-time Markov chain.

However, positive recurrence and null recurrence do not transfer easily to the continuous-time Markov chain.

In a continuous-time Markov chain, there's no notion of periodicity, since we are dealing with rates and not transition probabilities; there will always be a little probability flow going through each edge at each step.

Define P_t as the matrix given by $P_t(i,j) = \mathbb{P}[X_t = j \mid X_0 = i]$. Note that, because of memorylessness, $P_{t+s} = P_t P_s$.

Theorem 194 (Kolmogorov Equations). We have that

$$\frac{\mathrm{d}}{\mathrm{d}t} P_t = Q P_t$$

and under some technical conditions,

$$\frac{\mathrm{d}}{\mathrm{d}t} P_t = P_t Q$$

In other words, sometimes P_t and Q commute.

Proof. A little technical. But, taking finite differences and limits help. □

Theorem 195 (Matrix Definition). By definition of P_t,

$$P_t = \mathrm{e}^{tQ} = \sum_{k=0}^{\infty} \frac{t^k Q^k}{k!} = \mathrm{e}^{Qt}$$

where the index variable k represents how many jumps are taken in the time t.

It follows that

$$\pi_t = \pi_0 \mathrm{e}^{Qt}$$

for the continuous-time Markov chain.

Theorem 196. We have that for all t,

$$\pi Q = 0 \quad \text{if and only if} \quad \pi P_t = \pi$$

This π is called the stationary or invariant distribution for the continuous-time Markov chain.

Definition 197. The balance equations follow from

$$\pi Q = 0$$

or, for all i,

$$\sum_j \pi(j) Q(j,i) = 0$$

or,

$$\sum_{j \neq i} \pi(j) Q(j,i) = -\pi(i) Q(i,i) = \pi(i) \sum_{j \neq i} Q(i,j)$$

which is analogous to the discrete-time probability flow equations.

Theorem 198 (Ergodic Theorem for Continuous-Time Markov Chains). Consider an irreducible continuous-time Markov chain $(X_t)_{t \geq 0}$.

- If $(X_t)_{t \geq 0}$, is positive recurrent, then there exists a unique invariant and stationary distribution. If there exists an invariant distribution, then $(X_t)_{t \geq 0}$ is positive recurrent and the invariant distribution is unique.

- If $(X_t)_{t\geq 0}$ is positive recurrent, then
$$\mathbb{P}\left[\lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbb{1}(X_u = i)\,\mathrm{d}u = \pi(i)\right] = 1$$
In other words, if $Y_t(i) = \frac{1}{t}\int_0^t \mathbb{1}(X_u = i)\,\mathrm{d}u$, then $Y_t(i) \overset{\text{a.s.}}{\to} \pi(i)$. This is the ergodicity condition.
- If $(X_t)_{t\geq 0}$ is positive recurrent, then $\lim_{t\to\infty} P_t(i,j) = \pi(i)$.
- If $(X_t)_{t\geq 0}$ is not positive recurrent, then there is no invariant distribution and for all i we have $\lim_{t\to\infty} Y_t(i) \overset{\text{a.s.}}{\to} 0$.

Consider a continuous-time Markov chain $(X_t)_{t\geq 0}$ with embedded discrete-time Markov chain $(Y_n)_{n\in\mathbb{N}}$. Assume that both are irreducible and positive recurrent. Let π and α be the corresponding stationary distributions. Then
$$\pi(i) = \frac{\alpha(i)/q(i)}{\sum_k \alpha(k)/q(k)}$$
The intuition for this being true is that $\alpha(i)$ is the fraction of visits spent in state i, and at each visit we spend in expectation $1/q(i)$ amount of time. Thus we spend $\alpha(i)/q(i)$ time in state i. The denominator is just the normalization constant.

Example 199 (On-Off Source). Let state 0 be the state that a switch is off, and state 1 be the state that a switch is on. Let $Q = \begin{bmatrix}-\lambda & \lambda \\ \mu & -\mu\end{bmatrix}$. Then $\lambda\pi(0) = \mu\pi(1)$, so $\lambda(1) = \frac{\lambda}{\mu}\pi(0)$. Since $\pi(0)+\pi(1) = 1$, then $\frac{\lambda}{\mu}\pi(0)+\pi(0) = 1$, so $\pi(0) = \frac{\mu}{\lambda+\mu}$, and $\pi(1) = \frac{\lambda}{\lambda+\mu}$.

Example 200. Consider the queue example. If $\lambda > \mu$, then the probability mass in the chain drifts to ∞, implying that the chain is transient. If $\lambda = \mu$, then the probability mass stays stationary on average, but the expectation of return time to a state is ∞, implying that the chain is null recurrent. If $\lambda < \mu$, then the probability mass starting at k stays in the long run in the states in the interval $[1,k]$, and so the chain is positive recurrent. The invariant distribution, if it exists, is $\pi(i) = \left(\frac{\mu-\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^i$.

Definition 201. The detailed balance equations also exist for continuous-time Markov chains. Indeed, the detailed balance equations are:
$$\pi(i)Q(i,j) = \pi(j)Q(j,i)$$
The cut property also holds, that is, for each $S \subseteq \mathcal{X}$,
$$\sum_{(x,y)\in S\times(V\setminus S)} \pi(x)Q(x,y) = \sum_{(y,x)\in(V\setminus S)\times S} \pi(y)Q(y,x)$$
Recall that
$$\pi(i) = \frac{\alpha(i)/q(i)}{\sum_k \alpha(k)/q(k)}$$
If each $q(i) = q$ for some q, then $\pi(i) = \alpha(i)$. Note that by definition of the embedded discrete-time Markov chain, $X_t = Y_{N_t}$, for N_t the Poisson process which has an arrival whenever X_t transitions. Define $q = \sup\{q_i\}$ and $R = I + Q/q$. Then R is a legitimate probability transition matrix; $\pi R = \pi$ if and only if $\pi Q = 0$. Let $(Z_n)_{n\in\mathbb{N}}$ be the discrete-time Markov chain with matrix R. Then $(Z_n)_{n\in\mathbb{N}} = Y_{N'_t}$, where N'_t is a Poisson process with parameter q, and π is the stationary distribution of $(Z_n)_{n\in\mathbb{N}}$.

Theorem 202 (Poisson Arrivals See Time Averages). An arrival from a Poisson process observes the system as if it was arriving at a random moment in time. Therefore, the expected value of any parameter of the queue at the instant of a Poisson arrival is simply the long-run average value of that parameter.

Example 203. Consider the queue example, and put $\lambda < \mu$. The number of people waiting at any given time is the states, and $\pi(i) = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^i$. By PASTA, the total delay D (waiting and service), is computed. Indeed, $M_D(s) = \mathbb{E}\left[\mathrm{e}^{sD}\right] = \mathbb{E}\left[\mathbb{E}\left[\mathrm{e}^{sD} \mid T_S\right]\right] = \sum_{i=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^i\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\mu}{\mu-s}\right)^{i+1} = \frac{\mu-\lambda}{\mu-\lambda-s}$, so $D \sim \text{Exponential}(\mu-\lambda)$.

Now consider the hitting time equations. Let $T_A = \inf(\{t \geq 0 \colon X_t \in A\})$ and $\beta(i) = \mathbb{E}[T_A \mid X_0 = i]$. Then for each $i \in A$ we have $\beta(i) = 0$ and for $i \notin A$ we have

$$\beta(i) = \frac{1}{q(i)} + \sum_{j \neq i} \frac{Q(i,j)}{q(j)} \beta(j)$$

The hitting probability equations are similar. Let $\alpha(i) = \mathbb{P}[T_A < T_B \mid X_0 = i]$, for $A \cap B = \emptyset$. Then for each $i \in A$ we have $\alpha(i) = 1$, for each $i \in B$ we have $\alpha(i) = 0$, and otherwise

$$\alpha(i) = \sum_{j \neq i} \frac{Q(i,j)}{q(i)} \alpha(j)$$

Erdös-Rényi Random Graphs

Given an $n \in \mathbb{N}$, and $p \in [0,1]$, we define a graph distribution $G(n,p)$ is a random undirected graph with n vertices, and each of $\binom{n}{2}$ edges occur with probability p. There are $2^{\binom{n}{2}}$ such graphs.

If $p = 0$, then we get the empty graph on n vertices. If $p = 1$, then we get the complete graph. The question to discuss is the limiting behavior in n.

Let G_0 be sampled from $G(n,p)$. Then $\mathbb{P}[g \in G(n,p) = G_0] = p^{|E(g)|}(1-p)^{\binom{n}{2}-|E(g)|}$ where m is the number of edges of g. The expected number of edges in $G(n,p)$ is $\mathbb{E}[|E(G(n,p))|] = p\binom{n}{2}$. The degree of an arbitrary vertex is $\mathbb{P}[d(v) = d] = \binom{n-1}{d} p^d (1-p)^{n-1-d}$. The expected value of the degree is $\mathbb{E}[d(v)] = (n-1)p$.

We're concerned about the properties of $G(n,p(n))$, that is, p is a function of n. Define $p(n) = \frac{\lambda}{n}$, for $\lambda \to 0$. Then $\mathbb{E}[d(v)] = \frac{n-1}{n}\lambda$, and $\lim_{n\to\infty} \mathbb{E}[d(v)] = \lambda$, so $d(v) \sim \text{Poisson}(\lambda)$. The probability that a given vertex is disconnected is $\mathbb{P}[d(v) = 0] = (1-p)^{n-1}$ and in the limit this becomes $\mathrm{e}^{-\lambda}$.

Theorem 204 (Erdös-Renyi Theorem). Let $p(n) = \lambda \cdot \frac{\log(n)}{n}$, for $\lambda = 0$. If $\lambda < 1$, then $\lim_{n\to\infty} \mathbb{P}[G(n,p) \text{ is connected}] = 0$. If $\lambda > 1$, then $\lim_{n\to\infty} \mathbb{P}[G(n,p) \text{ is connected}] = 1$.

Proof. Let X_n be the number of isolated vertices in some graph $G(n, \lambda \cdot \frac{\log(n)}{n})$, where $\lambda \neq 1$. We want to show that $\lim_{n\to\infty} \mathbb{P}[X_n > 0] = \mathbb{1}(\lambda < 1)$.

Consider the case $\lambda < 1$. Then

$$\begin{aligned}
\mathbb{P}[X_n = 0] &= \mathbb{P}[|X_n - \mathbb{E}[X_n]| - \mathbb{E}[X_n]] \\
&\leq \mathbb{P}[|X_n - \mathbb{E}[X_n]|] \\
&\leq \frac{\operatorname{Var}[X_n]}{\mathbb{E}[X_n]^2} && \text{(Chebyshev inequality.)} \\
&= \frac{n \cdot \operatorname{Var}[Y_i] + n(n-1) \cdot \operatorname{Cov}[Y_i, Y_j]}{n^2 \cdot \mathbb{E}[Y_i]^2} && (Y_i = \mathbb{1}\left(i^{\text{th}} \text{ vertex is isolated}\right))
\end{aligned}$$

The evaluation of these quantities is tedious and left to the reader, as is the second part of the proof, which goes the same way. □

A result of this is that two vertices in a percolation network are connected almost surely if $p > \frac{1}{2}$, and two vertices are disconnected almost surely if $p < \frac{1}{2}$.

6 Estimation and Hypothesis Testing

We work with two kinds of estimators: maximum likelihood estimators, and maximum a posteriori estimators. The first utilizes Bayesian statistics, and the second uses probability theory and classical statistics.

In the Bayesian sense, let X be generated by a process with parameter Θ. Then Bayes' rule states that

$$p_{\Theta|X}(\theta \mid X) = \frac{p_{X|\Theta}(x \mid \theta)p_\Theta(\theta)}{p_X(x)}$$

In this case, the **prior** is $p_\Theta(\theta)$, the prior belief of the distribution of the parameter. The **posterior** probability is $p_{\Theta|X}$; the maximum a posteriori estimator is given by

$$\theta_{\text{MAP}} = \operatorname*{argmax}_\theta p_{\Theta|X}(\theta \mid x) = \operatorname*{argmax}_\theta \frac{p_{X|\Theta}(x \mid \theta)p_\Theta(\theta)}{p_X(x)} = \operatorname*{argmax}_\theta p_{X|\Theta}(x \mid \theta)p_\Theta(\theta)$$

On the other hand, the maximum likelihood estimator is simply given by

$$\theta_{\text{MLE}} = \operatorname*{argmax}_\theta p_{X|\Theta}(x \mid \theta)$$

This easily shows that if $p_\Theta(\theta) = c$ for some $c \in \mathbb{R}$, then $\theta_{\text{MLE}} = \theta_{\text{MAP}}$.

Example 205. Suppose you pick a coin at random and toss it n times, obtaining k heads. Then the maximum likelihood estimator is $p_{\text{MLE}} = \frac{k}{n}$.

At this point it's good to note that another way to write $\theta_{\text{MLE}} = \text{MLE}(\theta \mid x) = \operatorname{argmax}_\theta p_{X|\Theta}(x \mid \theta)$ and $\theta_{\text{MAP}} = \text{MAP}(\theta \mid x) = \operatorname{argmax}_\theta p_{\Theta|X}(\theta \mid x)$.

Example 206 (Romeo and Juliet). Juliet is late by some random time $x \sim \text{Uniform}([0, \theta])$ and $\theta \sim \text{Uniform}([0, 1])$. We attempt to compute $\text{MAP}(\theta \mid x)$. Since $p_\Theta(\theta) = \mathbb{1}(0 \le \theta \le 1)$ and $p_{X|\Theta}(x \mid \theta) = \frac{1}{\theta} \cdot \mathbb{1}(0 \le x \le \theta)$, we have

$$\begin{aligned}
\text{MAP}(\theta \mid x) &= \operatorname*{argmax}_\theta p_{\Theta|X}(\theta \mid x) \\
&= \operatorname*{argmax}_\theta p_{X|\Theta}(x \mid \theta)p_\Theta(\theta) \\
&= \operatorname*{argmax}_\theta \frac{\mathbb{1}(0 \le x \le \theta \le 1)}{\theta} \\
&= x
\end{aligned}$$

Thus, $\text{MAP}(\theta \mid x) = x$. Since Θ is uniform, $\text{MLE}(\theta \mid x) = x$ as well.

Example 207 (Binary Symmetric Channel). Assume that the input x to a binary symmetric channel with error probability $p \in (0, 1)$ is 1 with probability α and 0 with probability $1 - \alpha$. Then we estimate the input given $y = 1$.

$$\mathbb{P}[x = 0 \mid y = 1] = \frac{\mathbb{P}[y = 1 \mid x = 0]\mathbb{P}[x = 0]}{\mathbb{P}[y = 1]} = \frac{p(1 - \alpha)}{\mathbb{P}[y = 1]}$$

Similarly,

$$\mathbb{P}[x = 1 \mid y = 1] = \frac{\mathbb{P}[y = 1 \mid x = 1]\mathbb{P}[x = 1]}{\mathbb{P}[y = 1]} = \frac{(1 - p)\alpha}{\mathbb{P}[y = 1]}$$

Thus, $\text{MAP}(x \mid y = 1) = \mathbb{1}(p \le \alpha)$. Similarly, $\text{MAP}(x \mid y = 0) = \mathbb{1}(\alpha > 1 - p)$.

Example 208 (Gaussian Channel). Let $Z \sim \text{Normal}(0, \sigma^2)$. We send $X \in \{0, 1\}$, and the response is $y = x + z$. We decide whether $X = 0$ or $X = 1$, given Y. We maintain that $X \sim \text{Bernoulli}(\alpha)$. Note that $y \mid x \sim \text{Normal}(x, \sigma^2)$, from which we can derive the maximum likelihood estimator. Then

$$p_{X|Y}(x \mid y) = \frac{p_{Y|X}(y \mid x)p_X(x)}{p_Y(y)}$$

Finding the maximum a posteriori estimate does not require computing $\mathbb{P}[y]$, so plugging these in implies we're done. We obtain $\text{MAP}(x \mid y) = \mathbb{1}\left((1 - \alpha)p_{Y|X}(y \mid 0) \le \alpha p_{Y|X}(y \mid 1)\right) = \mathbb{1}\left(y \ge \frac{1}{2} + \sigma^2 \log\left(\frac{1-\alpha}{\alpha}\right)\right)$ and $\text{MLE}(x \mid y) = \mathbb{1}\left(p_{Y|X}(y \mid 1) > p_{Y|X}(y \mid 0)\right)$.

Suppose we have some variable $\theta \in \{0, 1\}$ and observe some output X, where $p_{X|\theta}$ is known. We wish to determine $\hat{\theta}$ which maximizes $\text{PCD} = \mathbb{P}\left[\hat{\theta} = 1 \mid \theta = 1\right]$ subject to $\text{PFA} = \mathbb{P}\left[\hat{\theta} \mid \theta = 0\right] \le \beta$. If the solution is a function R such that $\text{PCD} = R(\beta)$, then $R(\beta)$ is called the Receiver Operating Characteristic (ROC).

Theorem 209 (Neyman-Pearson). Define $L(x) = \frac{p_{X|\theta}(x|1)}{p_{X|\theta}(x|0)}$. The decision $\hat{\theta}$ that maximizes the PCD subject to PFA $\leq \beta$ is given by

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

where $\lambda > 0$ and $\gamma \in [0, 1]$ are chosen such that $\mathbb{P}\left[\hat{\theta} = 1 \mid \theta = 0\right] = \beta$.

Proof. Consider another decision rule $\widetilde{\Theta}$ such that

$$\mathbb{P}\left[\widetilde{\Theta} = 1 \mid \Theta = 0\right] \leq \beta$$

We need to show that

$$\mathbb{P}\left[\widetilde{\Theta} = 1 \mid \Theta = 1\right] \leq \mathbb{P}\left[\hat{\Theta} = 1 \mid \Theta = 1\right]$$

Observe

$$\left(\hat{\Theta} - \widetilde{\Theta}\right)(L(X) - \lambda) \geq 0$$

Taking expectations given $\Theta = 0$, we have

$$\mathbb{E}\left[\hat{\Theta}L(X) \mid \Theta = 0\right] - \mathbb{E}\left[\widetilde{\Theta}L(X) \mid \Theta = 0\right] \geq \lambda\left(\mathbb{E}\left[\hat{\Theta} \mid \Theta = 0\right] - \mathbb{E}\left[\widetilde{\Theta} \mid \Theta = 0\right]\right) = \lambda(\beta - \mathbb{P}\left[\widetilde{\Theta} = 1 \mid \Theta = 0\right])$$

Thus

$$\mathbb{E}\left[\hat{\Theta}L(X) \mid \Theta = 0\right] \geq \mathbb{E}\left[\widetilde{\Theta}L(X) \mid \Theta = 0\right]$$

By computation we obtain

$$\mathbb{E}[g(X)L(X) \mid \Theta = 0] = \mathbb{E}[g(X) \mid \Theta = 1]$$

Thus if $X = g(X, Z)$, where Z is independent of X and Θ, then

$$\mathbb{E}\left[\hat{\Theta}L(X) \mid \Theta = 0\right] = \mathbb{E}\left[\hat{\Theta} \mid \Theta = 1\right] = \mathbb{P}\left[\hat{\Theta} \mid \Theta = 1\right] \geq \mathbb{E}\left[\widetilde{\Theta} \mid \Theta = 1\right] = \mathbb{P}\left[\widetilde{\Theta} = 1 \mid X = 1\right]$$

as desired. □

If $L(x)$ is large, then for the observed x the outcome that $\theta = 1$ is more likely. As λ decreases, we choose $\hat{\theta} = 1$ more frequently; this causes PCD and PFA to increase.

Example 210. Recall the Gaussian channel, where $x \sim \text{Bernoulli}(\alpha)$, $z \sim \text{Normal}\left(0, \sigma^2\right)$, and $y = x + z$. Recall that $\text{MLE}(x \mid y) = \mathbb{1}\left(y \geq \frac{1}{2}\right)$ and $\text{MAP}(x \mid y) = \mathbb{1}\left(y \geq \frac{1}{2} + \sigma^2 \log\left(\frac{1-\alpha}{\alpha}\right)\right)$. Then

$$L(y) = \frac{\mathbb{P}[y \mid x = 1]}{\mathbb{P}[y \mid x = 0]} = \frac{\exp\left(-\frac{(y-1)^2}{2\sigma^2}\right)}{\exp\left(-\frac{y^2}{\sigma^2}\right)} = \exp\left(\frac{2y - 1}{2\sigma^2}\right)$$

Note that $L(y)$ is strictly increasing in y, and continuous, so $\mathbb{P}[L(y) = \lambda] = 0$. Then there exists some y_0 such that

$$\hat{x} = \mathbb{1}(y \geq y_0)$$

It remains to find y_0. We wish to choose y_0 such that PFA $= \beta$, so $\mathbb{P}[\hat{x} = 1 \mid x = 0] = \mathbb{P}[y \geq y_0 \mid x = 0] = \beta$. This means that $\mathbb{P}\left[\text{Normal}\left(0, \sigma^2\right) \geq y_0\right] = \beta$, so $\mathbb{P}\left[\text{Normal}(0, 1) \geq \frac{y_0}{\sigma}\right] = \beta$. Define $y(\beta) = \frac{y_0}{\sigma}$ such that $\mathbb{P}[\text{Normal}(0, 1) \geq y(\beta)] = \beta$. Then $y_0 = y(\beta) \cdot \sigma$.

Similarly, $\text{PCD} = \mathbb{P}[\hat{x} = 1 \mid x = 1] = \mathbb{P}\left[\text{Normal}\left(1, \sigma^2\right) \geq y_0\right] = \mathbb{P}\left[\text{Normal}(0, 1) \geq y(\beta) - \frac{1}{\sigma}\right]$.

Note that $y(\beta)$ is the ROC curve.

The perfect ROC curve is the step function, since we don't want to get anything wrong i.e. $\mathbb{P}\left[\hat{\theta} = 1 \mid \theta = 0\right] = 0$. Less optimal curves look like sigmoids, almost, but converge to the line $\text{PCD} = \beta/\text{PFA}$.

Example 211. Two machines produce light bulbs, with lifespans distributed according to Exponential(λ_0) or Exponential(λ_1) respectively. Without loss of generality, $\lambda_0 < \lambda_1$. Let $y = (y_1, \dots, y_n)$ be the lifespans observed from one machine in particular. We want to find which machine produced them.

We apply the Neyman-Pearson theorem. Let

$$L(y) = \frac{\mathbb{P}[y \mid x = 1]}{\mathbb{P}[y \mid x = 0]} = \frac{\prod_{i=1}^n \lambda_1 \exp(-\lambda_1 y_i)}{\prod_{i=1}^n \lambda_0 \exp(-\lambda_0 y_i)} = \left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\left(-(\lambda_1 - \lambda_0) \sum_{i=1}^n y_i\right)$$

Since $\lambda_1 > \lambda_0$, $L(y)$ is decreasing in $\sum_{i=1}^n y_i$. Since $L(y)$ is continuous, $\mathbb{P}[L(y) = \lambda] = 0$. Thus there exists some a such that $\hat{x} = \mathbb{1}(\sum_{i=1}^n y_i \le a)$. We choose a such that $\mathbb{P}[|sum_{i=1}^n y_i \le a \mid x = 0] = \beta$ Applying central limit theorem, we have $\sum_{i=1}^n y_i \le a$ if and only if $\frac{\sum_{i=1}^n y_i - \frac{n}{\lambda_0}}{\sqrt{n}} \le \frac{a - \frac{n}{\lambda_0}}{\sqrt{n}}$. For some confidence level β, we must have

$$\mathbb{P}\left[\text{Normal}(0,1) \le \lambda_0 \frac{a - \frac{n}{\lambda_0}}{\sqrt{n}}\right] = \beta$$

if z_β is the solution to $\int_{-z_\beta}^{z_\beta} p_{\text{Normal}(0,1)}(x)\,\mathrm{d}x = \beta$ then $a = \frac{a + z_\beta \sqrt{n}}{\lambda_0}$ and from this we can obtain the ROC curve and so on.

Example 212 (Randomization). Let $x \in \{0, 1\}$ and $y \in \{a, b, c\}$, with the following probability law: $\mathbb{P}[y = a \mid x = 1] = \frac{1}{5}$, $\mathbb{P}[y = b \mid x = 1] = \frac{1}{5}$, and $\mathbb{P}[y = c \mid x = 1] = \frac{3}{5}$. Similarly, $\mathbb{P}[y = a \mid x = 0] = \frac{1}{5}$, $\mathbb{P}[y = b \mid x = 0] = \frac{1}{2}$, and $\mathbb{P}[y = c \mid x = 0] = \frac{3}{10}$. Thus, $L(y)$ is determined by $L(a) = 1$, $L(b) = \frac{2}{5}$, and $L(c) = 2$.

For $\lambda = \frac{21}{10}$, we have PCD = PFA = 0. For $\lambda = 2$, we have PCD = $\frac{3}{5}\gamma$ and PFA = $\frac{3}{10}\gamma$. For $\lambda = 1$, we have PCD = $\frac{3}{5} + \frac{1}{5}\gamma$ and PFA = $\frac{3}{10} + \frac{1}{5}\gamma$. Note that varying $\gamma \in [0, 1]$, we can get a range of different PFA.

We start now with the linear least squares estimator.

The vector space axioms are written here for convenience. If V is a vector space over a field $\mathbb{F}$, then

- $\forall u \in V\ \forall v \in V\ (u + v \in V)$
- $\forall u \in V\ \forall v \in V\ \forall w \in V (u + (v + w) = (u + v) + w)$
- $\forall u \in V\ \forall v \in V\ (u + v = v + u)$
- $\exists 0 \in V\ \forall u \in V\ (u + 0 = u) \wedge \forall u \in V\ \exists - u \in V (u + (-u) = 0)$
- $\forall a \in \mathbb{F}\ \forall u \in V\ (au \in V)$
- $\exists 1 \in \mathbb{F}\ \forall u \in V\ (1u = u)$
- $\forall a \in \mathbb{F}\ \forall b \in \mathbb{F}\ \forall u \in V\ (a(bu) = (ab)u)$
- $\forall a \in \mathbb{F}\ \forall u \in V\ \forall v \in V\ (a(u + v) = au + av)$
- $\forall a \in \mathbb{F}\ \forall b \in \mathbb{F}\ \forall u \in V\ ((a + b)u = au + bu)$

Let $S \subseteq V$. Then $\text{span}(S) = \{\sum_{i=1}^n c_i v_i \colon n \in \mathbb{N}, v_i \in S, c_i \in \mathbb{F}\}$. We define $U \subseteq V$ as a subspace of V if U is itself a vector space. It's easy to show that for any $S \subseteq V$ that $\text{span}(S)$ is a subspace. S is linearly independent if and only there exists v such that there does not exist $n \in \mathbb{N}$ and v_i, c_i such that $\sum_{i=1}^n c_i v_i = v$. If S is linearly independent and $\text{span}(S) = V$, then S is a basis for V. It follows from the axiom of choice that for each vector space V, there exists a basis (possibly Hamel), and the cardinality of each basis is the same. Let $\dim(V)$ be the cardinality of this basis.

Vector spaces can also be inner product spaces. An inner product space is equipped with a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$ satisfying, for each $u, v, w \in V$ and $c \in \mathbb{F}$, that $\langle u, v \rangle = \langle v, u \rangle$, $\langle u + cv, w \rangle = \langle u, w \rangle + c \langle v, w \rangle$ (and indeed $\langle \cdot, \cdot \rangle$ is bilinear), and $\langle u, u \rangle \ge 0$ with equality if $u = 0$. We may equip such a space with some geometry. Define $\|\cdot\| \colon V \to [0, \infty)$ by $\|v\| = \sqrt{\langle v, v \rangle}$. We can also define the angle θ between u and v by the implicit definition $\langle u, v \rangle = \|u\| \|v\| \cos(\theta)$. From this, u and v are orthogonal if and only if $\langle u, v \rangle = 0$.

Such a space can also be a metric space if the function $d(u, v) \colon V \times V \to \mathbb{F}$ given by $d(u, v) = \|u - v\|$ obeys $d(u, v) \ge 0$ with equality only when $u = v$, $d(u, v) = d(v, u)$, and $d(u, v) \le d(u, w) + d(w, v)$.

We may also say such a space is complete if for each sequence $\{a_i\}_{i=1}^{\infty}$ of elements of V such that $\lim_{n\to\infty} a_{n+1}-a_n = 0$, we have $\lim_{n\to\infty} a_n \in V$. obeys

Definition 213 (Hilbert Space). A Hilbert space $\mathcal{H}$ is a complete inner product space over $\mathbb{C}$.

Given a fixed probability space Ω, define the Hilbert space of random variables $\mathcal{H} = \left\{X\colon \Omega \to \mathbb{R} \,\middle|\, \mathbb{E}\left[X^2\right] < \infty\right\}$. Define its inner product as $\langle X, Y\rangle = \mathbb{E}[XY]$. We confirm that $\mathcal{H}$ is an inner product space: since $X \in \mathcal{H}$ and $Y \in \mathcal{H}$, $X + Y \in \mathcal{H}$, so $\mathbb{E}[XY] \le \mathbb{E}\left[(X+Y)^2\right] = \mathbb{E}\left[X^2\right] + 2\mathbb{E}[XY] + \mathbb{E}\left[Y^2\right] \le \mathbb{E}\left[X^2\right] + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}\left[Y^2\right] < \infty$, so $\mathcal{H}$ is closed under addition. For finite or countable Ω, we have $\mathbb{E}[XY] = \sum_{\omega\in\Omega} X(\omega)Y(\omega)\mathbb{P}[\omega]$. If Ω is countable, then $\{\mathbb{1}(\omega)\}_{\omega\in\Omega}$ is a basis for $\mathcal{H}$.

Now we continue to estimation. The goal is to find $L(Y \mid X)$, which is the best linear estimate of Y given X. More abstractly, given $X \in V$ with subspace $U \subseteq V$, $L(Y \mid X)$ is the closest $X \in U$ to Y.

Given a subspace $U \subseteq V$, the set $U^{\perp} = \{v \in V\colon \langle u, v\rangle = 0 \ \forall u \in U\}$ is a subspace.

Define the orthogonal projection from V onto a subspace U as $p_U\colon V \to U$ defined by $p_U(y) = \operatorname{argmin}_{x\in U} \|y - x\|$. Since $p_U(y) \in U$, we have $y - p_U(y) \in U^{\perp}$. Let V have some countable orthonormal basis (where orthonormal means orthogonal and unit norm) $\{v_i\}_{i\in\mathbb{N}}$. Then $p_U(y) = \sum_{i=1}^{\infty} \langle y, v_i\rangle\, v_i$. Thus $c_i = \langle p_U(y), v_i\rangle = \langle y, v_i\rangle$.

Theorem 214 (Gram-Schmidt). The Gram-Schmidt Process converts any countable basis into an orthonormal one.

Let $\{v_i\}_{i\in\mathbb{N}}$ be a basis of some inner product space V, and let $\{u_i\}_{i\in\mathbb{N}}$ be defined initially by $v_i = u_i$. Let $S_n = \operatorname{span}(\{u_i\}_{i=1}^{n})$. Then iteratively set $u_i = \frac{v_i - p_{S_{i-1}}(v_i)}{\left\|v_i - p_{S_{i-1}}(v_i)\right\|}$. Then $\{u_i\}_{i=1}^{n}$ is an orthonormal basis of V.

Definition 215 (Linear Least Square Errror). Let $X, Y \in \mathcal{H}$. Given non-constant X, we define the linear least square error as

$$\operatorname{LLSE}(Y \mid X) = L(Y \mid X) = \min_{a,b\in\mathbb{F}} \mathbb{E}\left[(Y - a - bX)^2\right]$$

We can think about this in some different ways. First, the expectation is differentiable, so

$$\frac{\partial}{\partial a}\mathbb{E}\left[(Y-a-bX)^2\right] \overset{\text{set}}{=} \frac{\partial}{\partial b}\mathbb{E}[(Y-a-bX)] \overset{\text{set}}{=} 0$$

and solving for a_{OPT} and b_{OPT} in terms of X and Y gives the minimizing (a, b); then $L(Y \mid X) = \mathbb{E}\left[(Y - a_{\mathsf{OPT}} - b_{\mathsf{OPT}}X)^2\right]$.

Another way around this is that all $a + bX$ are elements of $U = \operatorname{span}(1, X)$. Thus, $Y - a_{\mathsf{OPT}} - b_{\mathsf{OPT}}X \in U^{\perp}$. Since $Y - a_{\mathsf{OPT}} - b_{\mathsf{OPT}}X \perp 1$, we have $\langle Y - a_{\mathsf{OPT}} - b_{\mathsf{OPT}}X, 1\rangle = \mathbb{E}[Y - a_{\mathsf{OPT}} - b_{\mathsf{OPT}}X] = 0$, and since $Y - a_{\mathsf{OPT}} - b_{\mathsf{OPT}}X \perp X$, we have $\langle Y - a_{\mathsf{OPT}} - b_{\mathsf{OPT}}X, X\rangle = \mathbb{E}[(Y - a_{\mathsf{OPT}} - b_{\mathsf{OPT}}X)X] = 0$, and this is useful.

Another, easier way is below:

Claim 216. Given Y, $X \in \mathcal{H}$, we have

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

Proof. We have that $\{1, X\}$ is a basis for $U \subseteq \mathcal{H}$. By Gram-Schmidt, $\left\{1, \frac{X-\mathbb{E}[X]}{\sqrt{\operatorname{Var}[X]}}\right\}$ is an orthonormal basis for U. This is because $u_1 = v_1 = 1$, and $u_2 = \frac{v_2 - \langle v_2, 1\rangle}{\|v_2 - \langle v_2, 1\rangle\|} = \frac{X - \mathbb{E}[X]}{\|X - \mathbb{E}[X]\|} = \frac{X-\mathbb{E}[X]}{\sqrt{\operatorname{Var}[X]}}$. Then

$$\begin{aligned} L(Y \mid X) &= \operatorname*{proj}_{U}(Y) \\ &= \operatorname*{proj}_{1}(Y) + \operatorname*{proj}_{X}(Y) \\ &= \langle Y, 1\rangle + \left\langle Y, \frac{X - \mathbb{E}[X]}{\sqrt{\operatorname{Var}[X]}}\right\rangle \frac{X - \mathbb{E}[X]}{\sqrt{\operatorname{Var}[X]}} \end{aligned}$$

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

□

Claim 217. The squared error for the linear least squares estimator is

$$\mathbb{E}\left[(Y - L(Y \mid X))^2\right] = \operatorname{Var}[Y] - \frac{\operatorname{Cov}[X,Y]^2}{\operatorname{Var}[X]}$$

Proof. Replace X with $X - \mathbb{E}[X]$ and Y with $Y - \mathbb{E}[Y]$. Then

$$L(Y \mid X) = \frac{\operatorname{Cov}[X,Y]}{\operatorname{Var}[X]}X$$

Let $Z = Y - \operatorname{proj}_U(Y)$. Then $\|Z\|^2 = \|Y\|^2 - \|L(X \mid Y)\|^2$ by vector addition. By computing inner products we obtain

$$\mathbb{E}\left[\|Z\|^2\right] = \mathbb{E}\left[\|Y\|^2\right] - \mathbb{E}\left[\|L(X \mid Y)\|^2\right] = \operatorname{Var}[Y] - \frac{\operatorname{Cov}[X,Y]^2}{\operatorname{Var}[X]}$$

This equation is not affected by a change to $\mathbb{E}[X]$ or $\mathbb{E}[Y]$, so the proof is complete. □

Example 218. Let $Y = aX + Z$, where $\mathbb{E}[X] = \mathbb{E}[Z] = 0$, and X and Z are independent. Then $L(X \mid Y) = \frac{\operatorname{Cov}[X,Y]}{\operatorname{Var}[Y]}Y$. We have $\operatorname{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X(aX+Z)] = a\mathbb{E}\left[X^2\right]$. We have $\operatorname{Var}[Y] = a^2\operatorname{Var}[X] + \operatorname{Var}[Z] = a^2\mathbb{E}\left[X^2\right] + \operatorname{Var}[Z]$. Then

$$L(X \mid Y) = \frac{a\mathbb{E}\left[X^2\right]}{a^2\mathbb{E}[X^2] + \mathbb{E}[Z^2]}Y = \frac{1/a}{1 + \operatorname{SNR}^{-1}}Y$$

where $\operatorname{SNR} = \frac{a^2\mathbb{E}\left[X^2\right]}{\mathbb{E}[Z^2]}$ is the signal to noise ratio. As $\operatorname{SNR} \to 0$, $L(X \mid Y) \to 0$. As $\operatorname{SNR} \to \infty$, $L(X \mid Y) \to \frac{Y}{a}$.

Example 219. Let $X = aY + bY^2$, where $Y \sim \operatorname{Uniform}([0,1])$. Then $L(X \mid Y) = \mathbb{E}[X] + \frac{\operatorname{Cov}[X,Y]}{\operatorname{Var}[Y]}(Y - \mathbb{E}[Y])$. We have $\mathbb{E}[X] = \alpha\mathbb{E}[Y] + \beta\mathbb{E}\left[Y^2\right] = \frac{\alpha}{2} + \frac{\beta}{3}$. Also $\operatorname{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}\left[\alpha Y^2 + \beta Y^3\right] - \left(\frac{\alpha}{2} + \frac{\beta}{3}\right)\cdot\frac{1}{2}$. Simplifying obtains $\operatorname{Cov}[X,Y] = \frac{\alpha}{3} + \frac{\beta}{4} - \frac{\alpha}{4} - \frac{\beta}{6} = \frac{\alpha+\beta}{12}$. Also, $\operatorname{Var}[Y] = \frac{1}{12}$. Thus $L(X \mid Y) = -\frac{\beta}{6} + (\alpha+\beta)Y$.

Suppose we observe i.i.d. (x_i, y_i). We want to find $g(X) = a + bX$ such that $\mathbb{E}\left[(Y - a - bX)^2\right]$ is minimized. For each $k \in \{1, \dots, n\}$, we find $a = a_{\mathsf{OPT}}$ and $b = b_{\mathsf{OPT}}$ such that

$$S_k = \sum_{j=1}^{k}(y_j - a - by_j)^2$$

is minimized, by setting $\frac{\partial}{\partial a}S_k = 0$ and $\frac{\partial}{\partial b}S_k = 0$. After some algebra, we obtain that after k data points have been fed in is

$$a_{\mathsf{OPT}}(k) + b_{\mathsf{OPT}}(k)X = \mathbb{E}_k[Y] + \frac{\operatorname{Cov}_k[X,Y]}{\operatorname{Var}_k[X]}(X - \mathbb{E}_k[X])$$

where

$$\mathbb{E}_k[X] = \frac{1}{k}\sum_{j=1}^{k}x_j, \quad \mathbb{E}_k[Y] = \frac{1}{k}\sum_{j=1}^{k}y_j$$

and

$$\operatorname*{Cov}_k[X,Y] = \left(\frac{1}{k}\sum_{j=1}^{k}x_jy_j\right) - \mathbb{E}_k[X]\mathbb{E}_k[Y], \quad \operatorname*{Var}_k[X] = \left(\frac{1}{k}\sum_{j=1}^{k}x_j^2\right) - \mathbb{E}_k[X]^2$$

The expectations are unbiased estimators, that is, $\mathbb{E}[\mathbb{E}_k[X]] = \mathbb{E}[X]$. The variance is biased, that is, $\mathbb{E}[\operatorname{Var}_k[X]] =$

$\frac{k-1}{k}\mathbb{E}[\mathrm{Var}[X]]$, and from this it is easy to see the sample variance

$$\sigma_k^2(X) = \frac{1}{k-1}\sum_{j=1}^{k}(x_i - \mathbb{E}_k[X])^2$$

where $\mathbb{E}\left[\mathbb{E}\left[\sigma_k^2(X)\right]\right] = \mathbb{E}[\mathrm{Var}[X]]$, and the sample covariance

$$\sigma_k(U,V) = \frac{1}{k-1}\sum_{j=1}^{k}(u_i - \mathbb{E}_k[u])(v_i - \mathbb{E}_k[v])$$

where $\mathbb{E}[\mathbb{E}[\sigma_k(U,V)]] = \mathrm{Cov}[U,V]$.

Now we contemplate the minimum mean square estimate. Assume we know the joint distribution of X and Y.

Definition 220. The minimum mean squared error is

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

We want to find $\mathrm{MMSE}(Y \mid X) = g(X)$ that minimizes $\mathbb{E}\left[(Y - g(X))^2\right]$. Let $\mathcal{G}(X) = \left\{f(X)\colon f \in L^2(\mathbb{R})\right\}$; it's easy to show that $\mathcal{G}(X)$ is a vector space. Then the minimizing $g(X) = \mathrm{proj}_{\mathcal{G}(X)}(Y)$. Contrast this to the linear least squares estimator, $L(Y \mid X) = \mathrm{proj}_{\mathrm{span}(1,X)}(Y)$.

Note that $\mathbb{E}[Y \mid X]$ is a random variable, so the minimum mean squared error is a random function of X.

We have that, under the appropriate (Lebesgue or counting) measure,

$$\mathbb{E}[Y \mid X = x] = \int_y y \cdot p_{Y|X}(y \mid x)\,\mathrm{d}y = \int_y y\frac{p_{X,Y}(x,y)}{p_X(x)}\,\mathrm{d}y$$

As this is an expectation, it has some properties:

- Linearity: $\mathbb{E}[aY + bZ \mid X] = a\mathbb{E}[Y \mid X] + b\mathbb{E}[Z \mid X]$
- Factorization: $\mathbb{E}[h(X)Y \mid X] = h(X)\mathbb{E}[Y \mid X]$
- Smoothing: $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$
- Independence: if X and Y independent then $\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$

Lemma 221. For any $\phi\colon \mathbb{R} \to \mathbb{R}$, $\mathbb{E}[(Y - \mathbb{E}[Y \mid X])\phi(X)] = 0$. Also, if for all ϕ we have $\mathbb{E}[(Y - g(X))\phi(X)] = 0$, then $g(x) = \mathbb{E}[Y \mid X = x]$.

Proof. We prove the first claim.

$$\begin{aligned}\mathbb{E}[(Y - \mathbb{E}[Y \mid X])\phi(X)] &= \mathbb{E}[Y\phi(X)] - \mathbb{E}[\phi(X)\mathbb{E}[Y \mid X]] \\ &= \mathbb{E}[Y\phi(X)] - \mathbb{E}[\mathbb{E}[\phi(X)Y \mid X]] \\ &= \mathbb{E}[Y\phi(X)] - \mathbb{E}[Y\phi(X)] \\ &= 0\end{aligned}$$

Now we prove the second claim.

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

Define $T_1 = g(X) - \mathbb{E}[Y \mid X]$, $T_2 = g(X) - Y$, and $T_3 = \mathbb{E}[Y \mid X] - Y$. Further let $\phi(x) = g(x) - \mathbb{E}[Y \mid X = x]$. Then $\mathbb{E}[T_1T_2] = 0$ by factoring and $\mathbb{E}[T_1T_3] = 0$ by $\mathbb{E}[\mathbb{E}[Y \mid X] - Y] = 0$, so $\mathbb{E}\left[(g(X) - \mathbb{E}[Y \mid X])^2\right] = 0$, so $g(X) = \mathbb{E}[Y \mid X]$. □

We can also interpret the minimum mean squared estimator geometrically. Then

$$\mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2\right] \le \mathbb{E}\left[(Y - h(X))^2\right], \quad \forall\, h \in \mathcal{G}$$

Geometrically, we have

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

By the Pythagorean theorem, we must have that $\langle Y-\mathbb{E}[Y \mid X], \mathbb{E}[Y \mid X]-h(X)\rangle = 0$ if and only if $h(X) = \mathbb{E}[Y \mid X]$, which show geometrically that $\text{MMSE}(Y \mid X) = \mathbb{E}[Y \mid X]$.

In general, $|Y - L(Y \mid X)| \geq |Y - \mathbb{E}[Y \mid X]|$, so the minimum mean square estimator is better than the linear least square estimator.

One instance where the linear least square estimator is equivalent to the minimum mean square estimator is when X and Y are jointly Gaussian, that is, if there exists $A \in \mathbb{R}^{2\times k}$ and $Z \in \mathcal{H}^k$, where $Z_i \overset{\text{i.i.d.}}{\sim} \text{Normal}(0,1)$, such that

$$\begin{bmatrix} X \\ Y \end{bmatrix} = AZ + \begin{bmatrix} \mathbb{E}[X] \\ \mathbb{E}[Y] \end{bmatrix}$$

If $X \in \mathcal{H}^n$ and $Y \in \mathcal{H}^m$ are random vectors and $\Sigma_{X,Y} = \mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])^\mathsf{T}\right]$, then

$$L(Y \mid X) = \mathbb{E}[Y] + \Sigma_{Y,X}\Sigma_{X,X}^{-1}(X-\mathbb{E}[X])$$

and the error vector length is expected to be

$$\mathbb{E}[Y - L(Y \mid X)] = \text{trace}\Big(\Sigma_{Y,Y} - \Sigma_{Y,X}\Sigma_{X,X}^{-1}\Sigma_{X,Y}\Big)$$

More generally, the random variables $Y_1, \dots, Y_n$ are jointly Gaussian if

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

where $A \in \mathbb{F}^{n\times k}$ and $Z \in \mathcal{H}^{k\times 1}$, where each $Z_i \overset{\text{i.i.d.}}{\sim} \text{Normal}(0,1)$. If this is the case, then Y has the auto-covariance matrix $\Sigma_{Y,Y} = AA^\mathsf{T}$, and we denote $Y \sim \text{Normal}(\mathbb{E}[Y], \Sigma_{Y,Y})$. Then if and only if $\Sigma_{Y,Y}$ is invertible we have

$$p_Y(y) = \frac{1}{(2\pi)^{n/2}\det(\Sigma_{Y,Y})^{1/2}}\exp\left(-\frac{1}{2}(y-\mathbb{E}[Y])^\mathsf{T}\Sigma_{Y,Y}^{-1}(y-\mathbb{E}[Y])\right)$$

It's easy to show that independent normal random variables are jointly Gaussian (just by putting a diagonal auto-covariance matrix). However, if the normal random variables are dependent, then they have to be dependent in a specific (linear) way.

Example 222. Let $Y_i \sim \text{Normal}\left(0, \sigma_i^2\right)$ be independent. Then $Y = AZ$, where $Z_i \overset{\text{i.i.d.}}{\sim} \text{Normal}(0,1)$, and $A = \text{diag}(\{\sigma_i\}_{i=1}^n)$. Then $\Sigma_{Y,Y} = \text{diag}\left(\left\{\sigma_i^2\right\}_{i=1}^n\right)$. Thus independent Gaussian variables are jointly Gaussian.

Example 223. Let Y_1 and Y_2 be defined by the density

$$p_Y(y_1,y_2) = \frac{\mathbb{1}(y_1y_2>0)}{\pi\sigma_1\sigma_2}\exp\left(-\frac{1}{2}\left(\frac{y_1^2}{\sigma_1^2}+\frac{y_2^2}{\sigma_2^2}\right)\right)$$

Integrating to find the marginal densities show that Y_1 and Y_2 are independent, but their relationship is clearly not globally linear.

Theorem 224. Jointly Gaussian random variables are independent if and only if they are uncorrelated.

Proof. It's easy to show that if components are independent then they are uncorrelated. If components are uncorrelated, then A and thus $\Sigma_{Y,Y}$ are diagonal matrices, and from this we obtain that the joint density factors. □

The level curves with respect to the density are generalized ellipsoids in $\mathbb{F}^n$. Let $L(Y) = (Y-\mathbb{E}[Y])^\mathsf{T}\Sigma_{Y,Y}^{-1}(Y-\mathbb{E}[Y])$. If $\Sigma_{Y,Y}$ is invertible then there are n distinct positive real eigenvalues of $\Sigma_{Y,Y}$ and thus of $\Sigma_{Y,Y}^{-1}$. Let $\Sigma_{Y,Y}^{-1} = U\Lambda U^\mathsf{T}$, where U has columns that are orthonormal eigenvectors corresponding to the n distinct positive real eigenvalues

$\lambda_1 > \cdots > \lambda_n > 0$ of $\Sigma_{Y,Y}^{-1}$, and let $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Define $\widetilde{y} = U^\mathsf{T}(y - \mathbb{E}[Y])$, so then $L(\widetilde{y}) = \sum_{i=1}^n \lambda_i \widetilde{y}_i^n$. Set $L(\widetilde{y}) = c$ to obtain a level curve, obviously dividing by $\prod_i \lambda_i$ gives an ellipsoid.

For jointly Gaussian random variables X and Y, we claim $L(X \mid Y) = \mathbb{E}[X \mid Y] = \operatorname{MMSE}(X \mid Y)$. We have shown that $X - L(X \mid Y)$ and Y are uncorrelated, and since they are linear functions of X and Y, we have $X - L(X \mid Y)$ and Y are independent. Further, for any square-integrable ϕ, $X - L(X \mid Y)$ and $\phi(Y)$ are independent. Then it's possible to take inner products and show the result.

We now begin to think about Kalman filtering. In particular, we discover how to update the linear least squares estimator with orthogonal observations. Let Y and $X_1, \ldots, X_n$ be zero-mean with X_i being pairwise orthogonal. Then

$$L(Y \mid X_1, \ldots, X_n) = \sum_{i=1}^n L(Y \mid X_i)$$

If Y is not zero-mean, then

$$L(Y \mid X_1, \ldots, X_n) = \mathbb{E}[Y] + \sum_{i=1}^n L(Y - \mathbb{E}[Y] \mid X_i) = \sum_{i=1}^n L(Y \mid X_i) - \mathbb{E}[Y]$$

We want to show that $X_i - (\sum_{j=1}^n L(Y \mid X_i))$ is orthogonal to each X_j. Indeed, rewrite this sum as $(X_i - L(Y \mid X_j)) - \sum_{k \neq i} L(Y \mid X_k)$. The first term is orthogonal to X_j by design; the second term is linear in all the X_k for $k \neq j$, and so orthogonal to X_j, so the whole expression is orthogonal to X_j for whichever i and j we pick. If the X_i are not orthogonal, then

$$L(Y \mid X_1, \ldots, X_n) = \sum_{i=1}^n L\left(Y \,\middle|\, X_i - \sum_{j=1}^{i-1} L(X_i \mid X_j)\right)$$

due to the Gram-Schmidt process. If $\mathbb{E}\left[X_i - \sum_{j=1}^{i-1} L(X_i \mid X_j)\right] = 0$ then X_i is orthogonal to X_j for $j < i$, so we can apply the previous result.

Before we begin talking about the Kalman filter, we have some useful lemmas.

Lemma 225. The following are true:

- If X and Y are independent, then X and Y are uncorrelated (i.e. $\operatorname{Cov}[X, Y] = 0$).
- If X and Y are uncorrelated and jointly Gaussian, then X and Y are independent.
- If X and Y are orthogonal and $\mathbb{E}[X] = 0$ or $\mathbb{E}[Y] = 0$, then X and Y are uncorrelated.
- If X and Y are uncorrelated or independent and $\mathbb{E}[X] = 0$ or $\mathbb{E}[Y] = 0$, then X and Y are orthogonal.

We discuss the scalar Kalman filter for now. We have a system with state X_n and output Y_n, that follows the following recurrence:

$$X_n = aX_{n-1} + V_n$$
$$Y_n = cX_n + W_n$$

where $\left\{X_0, \{V_n\}_{n \in \mathbb{N}}, \{W_n\}_{n \in \mathbb{N}}\right\}$ are independent and zero-mean. We assume that $|a| < 1$ (because otherwise the system will blow up) and $c = 1$ (because we can just rescale), and that these values are known to us. We wish to compute $\hat{X}_n = L(X_n \mid Y_1, \ldots, Y_n)$.

Recall that if X, Y, Z are zero-mean random variables, then define $\widetilde{Z} = Z - L(Z \mid Y)$. Then we have that $\widetilde{Z}$ is orthogonal to 1 and Y, so $\left\{1, Y, \widetilde{Z}\right\}$ is orthogonal, and so

$$L(X \mid Y, \widetilde{Z}) = L(X \mid Y) + L(X \mid \widetilde{Z}) = L(X \mid Y) + L(X \mid Z - L(Z, Y)) = L(X \mid Y, Z)$$

where we take into account that $L(A \mid B) = \operatorname{proj}_{\operatorname{span}(B)}(A)$.

Since it seems important, define $A - L(A \mid B)$ as the innovation in A.

Now define $\widetilde{Y}_n = Y_n - L(X \mid L(Y_n \mid Y_1, \dots, Y_{n-1}))$, and define $\hat{X}_{n|k} = L(X_n \mid Y_1, \dots, Y_k)$. We obtain

$$\begin{aligned}\hat{X}_{n|n} &= L(X_n \mid Y_1, \dots, Y_n) \\ &= L(X_n \mid Y_1, \dots, Y_{n-1}) + L(X \mid Y_n - L(Y_n \mid Y_1, \dots, Y_{n-1})) \\ &= L(X_n \mid Y_1, \dots, Y_{n-1}) + L(X \mid \widetilde{Y}_n) \\ &= \hat{X}_{n|n-1} + k_n \widetilde{Y}_n \end{aligned} \tag{1}$$

and

$$\begin{aligned}\hat{X}_{n|n-1} &= L(X_n \mid Y_1, \dots, Y_{n-1}) \\ &= L(aX_{n-1} + V_n \mid Y_1, \dots, Y_{n-1}) \\ &= aL(X_{n-1} \mid Y_1, \dots, Y_{n-1}) + L(V_n \mid Y_1, \dots, Y_{n-1}) \\ &= a\hat{X}_{n-1|n-1} \end{aligned} \tag{2}$$

Also,

$$\widetilde{Y}_n = Y_n - L(Y_n \mid Y_1, \dots, Y_{n-1}) \tag{3}$$

$$= Y_n - L(X_n + W_n \mid Y_1, \dots, Y_{n-1})$$

$$= Y_n - L(X_n \mid Y_1, \dots, Y_{n-1}) \tag{4}$$

$$= Y_n - \hat{X}_{n|n-1} \tag{5}$$

Putting this all together, we obtain

$$\hat{X}_{n|n} = a\hat{X}_{n-1|n-1} + k_n(Y_n - \hat{X}_{n|n-1}) \tag{1, 2, 5}$$

$$= a\hat{X}_{n-1|n-1} + k_n(Y_n - a\hat{X}_{n-1|n-1}) \tag{6}$$

Our goal is to find k_n geometrically. We also define the gain of a Kalman filter as k_n.

We visualize $\mathcal{H}$ as $\mathbb{R}^3$. We fix an origin O and an X_n. Then we place $\hat{X}_{n|n-1}$ such that $\widetilde{X}_{n|n-1} \perp X_n \hat{X}_{n|n-1}$. Now we place $\widetilde{Y}_n$ such that $\widetilde{Y}_n \perp \hat{X}_{n|n-1}$ (due to (3) and (4)), and Y_n such that $Y_n = X_n = \hat{X}_{n|n-1} + \widetilde{Y}_n$ (due to (5)). Then we place $k_n\widetilde{Y}_n$ as $\operatorname{proj}_{\operatorname{span}\left(\widetilde{Y}_n\right)}(X_n)$.

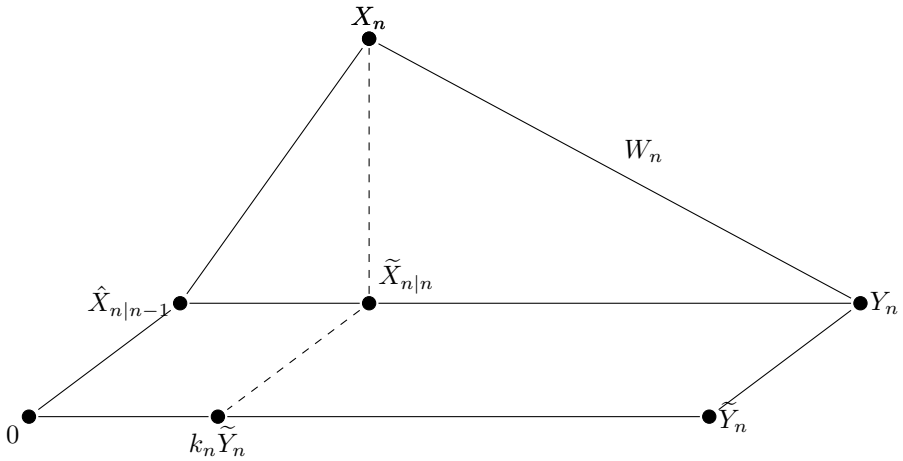


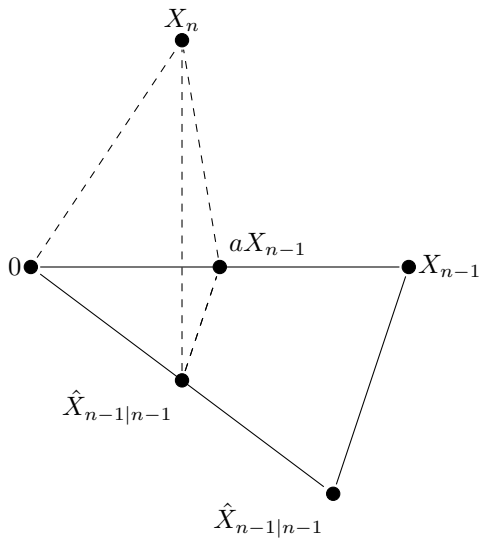
Figure 1: Initial geometry of the scalar Kalman figure.

Let $A = \hat{X}_{n|n-1}$, $B = \hat{X}_{n|n}$, $C = X_n$, and $D = Y_n$. Then $\triangle ABC \simeq \triangle ACD$. Also, $|AB| = k_n|AD|$, and $\frac{|AB|}{|AC|} = \frac{|AC|}{|AD|}$,

thus

$$k_n = \frac{|AC|^2}{|AD|^2} = \frac{|AC|^2}{|AC|^2 + |CD|^2} = \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_W}$$

We know σ^2_W, since it is the variance of noise. It remains to find $\sigma^2_{n|n-1}$. We also attempt to find this using geometry. Another facet of the above diagram is



Define $A = X_n$, $B = \hat{X}_{n|n-1}$, $C = aX_{n-1}$, $D = \hat{X}_{n-1|n-1}$, and $E = X_{n-1}$. Let $\Delta_{n|n-1} = \overline{\hat{X}_{n|n-1}X_n}$ and $\Delta_{n-1|n-1} = \overline{\hat{X}_{n-1|n-1}X_{n-1}}$. Then

$$\begin{aligned}\sigma^2_{n|n-1} &= \left|\Delta_{n|n-1}\right|^2 \\ &= |BC|^2 + |AC|^2 \\ &= |BC|^2 + \sigma^2_V\end{aligned}$$

and so $|BC| = a\left|\Delta_{n-1|n-1}\right|$.

We thus have

$$\sigma^2_{n|n-1} = a^2\sigma^2_{n-1|n-1} + \sigma^2_W$$

From Figure 1 we obtain

$$\begin{aligned}\sigma^2_{n|n} &= |BC|^2 = |AC|^2 - |AB|^2 \\ &= \sigma^2_{n|n-1} - k_n^2|AD|^2 \\ &= \sigma^2_{n|n-1} - k_n^2\frac{|AC|^2}{k_n} \\ &= \sigma^2_{n|n-1}(1 - k_n)\end{aligned}$$

For the purposes of the algorithm, we have

$$\hat{X}_{n|n} = a\hat{X}_{n-1|n-1} + k_n\left(Y_n - a\hat{X}_{n-1|n-1}\right)$$

(we compute $\hat{X}_{1|1}$ and iterate for $n \geq 2$), and

$$\sigma^2_{n|n-1} = a^2\sigma^2_{n-1|n-1} + \sigma^2_V$$

(we compute $\sigma^2_{1|1}$ and iterate for $n \geq 2$), and we have the equations

$$\sigma^2_{n|n} = \sigma^2_{n|n-1}(1 - k_n)$$

and

$$k_n = \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_W}$$

The vector case is decided similarly; that is, if

$$\begin{aligned} X_n &= AX_{n-1} + V_{n-1} \\ Y_n &= CX_n + W_n \end{aligned}$$

then we have

$$\begin{aligned} \hat{X}_{n|n} &= A\hat{X}_{n-1|n-1} + K_n\widetilde{Y}_n \\ \widetilde{Y}_n &= Y_n - CA\hat{X}_{n-1|n-1} \\ K_n &= \Sigma_{n|n-1}C^\mathsf{T}\left(C\Sigma_{n|n-1}C^\mathsf{T} + \Sigma_W\right)^{-1} \\ \Sigma_{n|n-1} &= A\Sigma_{n-1|n-1}A^\mathsf{T} + \Sigma_V \\ \Sigma_{n|n} &= (I - K_nC)\Sigma_{n|n-1} \end{aligned}$$

which mirrors the scalar case.