EECS 126 Lecture Notes

Alec Li

Spring 2021 — Professor Thomas Courtade

Contents

Theorems

Definitions

1/19/2021

Lecture 1

Axioms of Probability

Definition 1.1: Probability Space

A **probability space** is a triple $(\Omega, \mathcal{F}, P)$:

- Ω = "sample space" = set of "samples".
- $\mathcal{F}$ = family of subsets of Ω, called "events"
- P = probability measure

1.1 Axioms

- Technical assumption: $\mathcal{F}$ is a "σ-algebra" containing Ω itself.
  - a σ-algebra means that countable complements/unions/intersections of events are also events.
- $P: \mathcal{F} \to [0,1]$ assigns "probabilities" to events. Probability measures must obey Kolmogorov Axioms:

Definition 1.2: Kolmogorov Axioms

1. $P(A) \geq 0$, $\forall A \in \mathcal{F}$; probabilities are nonnegative
2. $P(\Omega) = 1$; probability of all possible outcomes is collectively equal to 1
   - Note: $\Omega \in \mathcal{F}$
3. If $A_1, A_2, \ldots \in \mathcal{F}$ and $A_i \cap A_j = \varnothing$, $\forall i \neq j$, then $P(\bigcup_{i\geq 1} A_i) = \sum_{i \geq 1} P(A_i)$; countable additivity: probabilities of disjoint events are additive
   - Why do we need countability? $1 = P([0,1]) = \sum_{x \in [0,1]} P(\{x\}) = 0$ because the probability of getting $x = k$ in a continuous distribution is 0.

Example 1.3

Flip a biased coin: heads with probability p, tails with probability $1-p$.

$$\Omega = \{H, T\}$$
$$\mathcal{F} = 2^{\Omega} = \{\varnothing, H, T, \{H, T\}\} \text{ (all possible subsets of } \Omega)$$
$$P(H) = p, P(T) = 1-p, P(\varnothing) = 0, P(\{H, T\}) = 1$$

$(\Omega, \mathcal{F}, P)$ is a valid probability space that describes an experiment.

The choice of the sample space and $\mathcal{F}$ can have flexibility; it is up to us to pick something that suitably models the problem.

Example 1.4

Same setup as in Example 1.3, but

$$\Omega = \{\text{all configurations of atoms in the universe}\}$$

And events

$$A = \{\text{set of configurations s.t. coin lands heads}\}$$
$$B = \{\text{set of configurations s.t. coin lands tails}\}$$

This means that $\mathcal{F} = \{\varnothing, A, B, \{A \cup B\} = \Omega\}$.

$(\Omega, \mathcal{F}, P)$ is another valid choice of probability space to model the problem.

Example 1.5

Flip two coins, biased to heads with probability p and q respectively.

$$\Omega = \{HH, HT, TH, TT\}$$
$$\mathcal{F} = \{\varnothing, \{HH, HT\}, \{TH, TT\}, \Omega\}$$
$$P(A) = p,\ P(B) = 1 - p$$

Or,

$$\mathcal{F} = 2^{\Omega}$$
$$P(HH) = pq,\ P(HT) = p(1-q),\ P(TH) = (1-p)q,\ P(TT) = (1-p)(1-q)$$

1.2 Consequences of the axioms

1. If Ω is countable, and each $\omega \in \Omega$ is an event, then $\sum_{\omega \in \Omega} P(\omega) = 1$.
2. $P(A^c) = 1 - P(A)$ for event A

   Proof.

   $$\begin{aligned} 1 &= P(\Omega) && \text{(Axiom 2)} \\ &= P(A \cup A^c) \\ &= P(A) + P(A^c) && \text{(Axiom 3)} \end{aligned}$$

   □

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

   Proof. If A, B are events, so are $A \cup B$, $A \cap B$.

   $$\begin{aligned} P(A \cup B) &= P(A \setminus (A \cap B)) + P(B) && \text{(Axiom 3)} \\ &= P(A) - P(A \cap B) + P(B) && \text{(Axiom 3)} \end{aligned}$$

   □

   More generally, inclusion-exclusion principle holds.

1/21/2021

Lecture 2

Conditional Probability, Independence

2.1 Conditional Probability

In applying probability, we often want to compute probabilities of "complicated" events.

Theorem 2.1: Law of Total Probability

If $A_1, A_2, \ldots$ partition Ω (i.e. A_i's are disjoint, and $\bigcup_{i\geq 1} A_i = \Omega$), then for any event B,

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

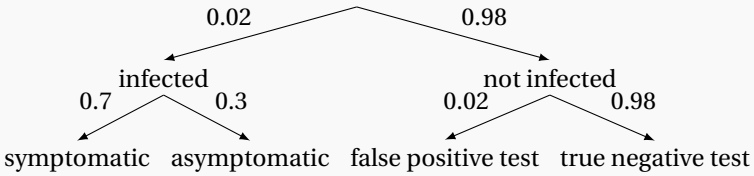
Proof. We can write

$$\begin{aligned}
\mathbb{P}(B) &= \mathbb{P}(B \cap \Omega) \\
&= \mathbb{P}\left(B \cap \bigcup_{i\geq 1} A_i\right) \\
&= \mathbb{P}\left(\bigcup_{i\geq 1} (A_i \cap B)\right) \\
&= \sum_{i\geq 1} \mathbb{P}(A_i \cap B) && \text{(Axiom 3)}
\end{aligned}$$

□

Example 2.2

Suppose you are part of surveillance testing for COVID. What's the probability that a random person in the surveillance pool tests positive (+) and is asymptomatic?



Here, we're assuming that all truly infected people will test positive.

This means that we have

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

Definition 2.3: Conditional Probability

If B is an event with $\mathbb{P}(B) > 0$, then the conditional probability of A given B is

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

If $\mathbb{P}(A_i) > 0$, then the Law of Total Probability becomes

$$\mathbb{P}(B) = \sum_{i\geq 1} \mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i).$$

Theorem 2.4: Bayes Rule

If events A, B have positive probability, then

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

Proof.

$$\begin{aligned}\frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)} &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \cdot \frac{\mathbb{P}(A)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\ &= \mathbb{P}(A \mid B) && \text{(definition of conditional probability)}\end{aligned}$$

□

Example 2.5: Conditional Probability

Suppose you are part of surveillance testing. Given that you test positive and are asymptomatic, what is the probability that you are actually infected?

$$\mathbb{P}\big(\text{infected} \mid (+, \text{asymptomatic})\big) = \frac{\mathbb{P}\big(\text{infected} \cap (+, \text{asymptomatic})\big)}{\mathbb{P}\big(+, \text{asymptomatic}\big)}$$

Example 2.6

We roll 2 standard die, and the sum is 10. What is the probability that the first roll was a 4?

Let A = first roll is 4, and B = sum of rolls is 10.

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

For conditional probabilities, we can more generally decompose (provided these conditional probabilities actually exist):

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

This basically decomposes the intersection into a sequence of events.

Example 2.7: Birthday Paradox

Let n people be in a room. What is the probability that no two people share a common birthday?

For $i = 1, \ldots, n$ and $j = 1, \ldots, i-1$, let $A_i = \{\text{person } i \text{ does not share a birthday with any person } j\}$.

Then,

$$\mathbb{P}\big(\text{no people share common birthday}\big) = \mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathbb{P}(A_i \mid A_1 \cap \cdots \cap A_{i-1}).$$

We can easily find $\mathbb{P}(A_i \mid A_1 \cap A_2 \cap \cdots \cap A_{i-1}) = \frac{365-(i-1)}{365}$, because we have $i-1$ less choices of days for each person i.

This means we have

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) &= \prod_{i=1}^{n}\left(1-\frac{i-1}{365}\right)\\ &\le \prod_{i=1}^{n}\exp\left(-\frac{i-1}{365}\right)\\ &= \exp\left(-\sum_{i=1}^{n}\frac{i-1}{365}\right)\\ &= \exp\left(-\frac{\binom{n}{2}}{365}\right)\end{aligned}$$

The inequality comes from $1-x \le e^{-x}$, and approximately equal when x is small.

So, we also have that $\mathbb{P}\big(2 \text{ people share common birthday}\big) \ge 1-\exp\left(-\frac{\binom{n}{2}}{365}\right)$. With $n=23$, that's a probability of about 0.5.

2.2 Independence

Definition 2.8: Independence

Events A, B are independent if

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

Note: Sometimes this "technical" definition of independence does not agree with intuition, but it usually does.

Note: If $\mathbb{P}(B) > 0$, then A, B independent iff $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. In other words, knowing B occurred tells us nothing about A.

In general, events $A_1, A_2, \cdots$ are independent if

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

for all finite subsets of indices S.

A fact that you can prove: if $A_1, A_2, \cdots$ are independent, then so are $B_1, B_2, \cdots$, where each B_i is equal to A_i or A_i^c. In other words, you can complement some events and they stay independent.

Example 2.9

Consider an infinite sequence of independent fair coin tosses.

What does this mean, precisely?

Let $A_i = \{\text{toss } i \text{ is heads}\}$; $A_i^c = \{\text{toss } i \text{ is tails}\}$. The probability of any given sequence of tosses is

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n}\{\text{toss } i \text{ has outcome } x_i\}\right) &= \prod_{i=1}^{n}\mathbb{P}(\{\text{toss } i \text{ lands } x_i\})\\ &= \prod_{i=1}^{n}\frac{1}{2} = 2^{-n}\end{aligned}$$

Definition 2.10: Conditional Independence

If events $A, B, C > 0$ with $P(C) > 0$ satisfy

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

then A and B are conditionally independent given C.

Example 2.11

Let $A = \{\text{random person is lactose intolerant}\}$ and $B = \{\text{random person is shorter than average}\}$.

A famous study showed that $\mathbb{P}(A \mid B) \gg \mathbb{P}(A)$; i.e. is lactose intolerance caused by height? No!

$C = \{\text{person is of Asian descent}\}$, then

$$\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid C) \iff \mathbb{P}(A \mid B \cap C) = \mathbb{P}(A \mid C).$$

2.3 Random Variables

Definition 2.12: Random Variable

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

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

The fact that this set is an event means that we can compute the probability

$$\mathbb{P}(X \leq \alpha) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \leq \alpha\}), \ \forall \alpha \in \mathbb{R}.$$

1/26/2021

Lecture 3

Random Variables, Expectation

3.1 Random Variables (aside)

The definition of random variables from last lectures ensures that we can compute $\mathbb{P}(X \in B)$ for pretty much any set B you'll be likely to encounter.

Why? Since $\mathcal{F}$ is closed under complements (by definition),

$$\{\omega \mid X(\omega) > \alpha\} \in \mathcal{F}, \ \forall \alpha \in \mathbb{R}.$$

Since $\mathcal{F}$ is closed under intersection, this implies that

$$\{\omega \mid \alpha < X(\omega) \leq \beta\} \in \mathcal{F}, \ \forall \alpha < \beta \in \mathbb{R}.$$

Since $\mathcal{F}$ is closed under countable unions, then this also implies that

$$\{\omega \mid \alpha < X(\omega) < \beta\} = \bigcup_{n \geq 1} \left\{\omega \mid \alpha < X(\omega) \leq \beta - \frac{1}{n}\right\},$$

and since the union of events is in $\mathcal{F}$, the LHS must also be in $\mathcal{F}$. (The RHS is approaching β but never equal to β.)

We can keep going, and as thus $\{X \in A\}$ for open set A is also an event, etc.

Aside: *Borel Sets* are sets that can be built up by unions/intersections/complements of open/closed sets.

From this seemingly basic definition of a random variable, we can use the properties of $\mathcal{F}$ and build from there, proving that we can compute the probability that X is in any complicated set that we're interested in.

This technical definition of a RV also implies that RVs satisfy certain algebraic properties:

- If X, Y are RVs, then so is $X+Y$, XY, X^p (for $p \in \mathbb{R}$)
- If $X_1, X_2, \cdots$ are RVs, then so is $\lim_{n\to\infty} X_n$ (if the limit exists)

3.2 Discrete Random Variables

Definition 3.1: Discrete Random Variable

A discrete RV is a RV that takes countably many values.

Some examples:

- X = roll of a die (takes values $\{1,\dots,6\}$)
- X = number of times I need to cast before I catch a fish (takes values $\{1,2,3,\cdots\}$)
- X = number of heads in n coin flips (takes values in $\{0,1,2,\dots,n\}$)
- X = fraction of heads in n coin flips (takes values in $\{0, \frac{1}{n}, \dots, \frac{n-1}{n}, 1\}$)

Definition 3.2: Probability Mass Function (pmf)

The frequencies with which a discrete RV takes different values are described by its *probability mass function* (pmf): $P_X(x) := \mathbb{P}(X = x) = \mathbb{P}(\{\omega \mid X(\omega) = x\})$

The PMF doesn't depend on individual samples ω; it's just the frequency by which the variable takes on values. The PMF is also defined as $P_X : \mathcal{X} \to [0,1]$.

For example

- If X is a fair coin flip where $H \to 1$ and $T \to 0$, then $P_X(1) = \frac{1}{2}$ and $P_X(0) = \frac{1}{2}$.

The PMF is also called the *distribution of* X.

Note: By Axiom 3,

$$\sum_{x\in\mathcal{X}} P_X(x) = 1.$$

Definition 3.3: Joint Probability Mass Function

If X, Y are RVs on a common probability space $(\Omega, \mathcal{F}, P)$, then their *joint pmf* describes frequencies of joint outcomes:

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

We can also infer several properties of joint distributions:

- By the law of total probability, we also see that

$$\sum_y P_{XY}(x,y) = P_X(x).$$

  This is called a marginal distribution.

- Random variables X, Y are independent if $P_{XY}(x,y) = P_X(x)P_Y(y)$.

  This is the same as saying the events $\{\omega \mid X(\omega) = x\}$ and $\{\omega \mid Y(\omega) = y\}$ are independent in $(\Omega, \mathcal{F}, P)$, $\forall x, y \in \mathcal{X} \times \mathcal{Y}$

Example 3.4

Suppose X_1, X_2 to be fair coin tosses, but linked via magic so that $X_1 = X_2$ with probability $\frac{3}{4}$. The pmf can be represented in this table:

		X_2	
		0	1
X_1	0	$\frac{3}{8}$	$\frac{1}{8}$
	1	$\frac{1}{8}$	$\frac{3}{8}$

where the columns are X_2 and the rows are X_1, i.e. $P_{X_1X_2}(x_1, x_2)$ is represented by this table.

Conceptual note: Often, it is most natural to model a problem in terms of random variables and their (joint) distributions, and no mention is made regarding the underlying probability space. This turns out to be okay!

Why? There is a deep theorem in probability (Kolmogorov Extension Theorem) which says that if random variables and their distributions are specified in a "consistent" way, then there exists an underlying probability space that gives rise to the desired joint distributions.

Explanation of notation: "Pr" to denote a generic probability assignment when the probability space has not been explicitly defined[1].

3.3 Expectation

Definition 3.5: Expectation

For a discrete RV X, the expectation of X is defined as

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

provided the series exists.

Theorem 3.6: Law of the Unconscious Statistician

If $Y = g(X)$, $g : \mathcal{X} \to \mathbb{R}$, then Y is a RV and

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

Proof.

$$\mathbb{E}[Y] = \sum_y y P_Y(y)$$

[1] This is not used in this set of lecture notes, but was used in lecture. I use $\mathbb{P}(X = x)$ for the most part, and sometimes $P_X(x)$ when referring to discrete probability mass functions—generally, the two are used interchangeably here.

$$
\begin{aligned}
&= \sum_y y \sum_{x|g(x)=y} P_X(x) \\
&= \sum_y \sum_{x|g(x)=y} g(x) P_X(x) \\
&= \sum_{x \in \mathcal{X}} g(x) P_X(x)
\end{aligned}
$$

□

One of the most important properties of expectation is the linearity of expectation:

Theorem 3.7: Linearity of Expectation

Expectation is linear, i.e.

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y], \ a, b \in \mathbb{R}.$$

Proof.

$$
\begin{aligned}
\mathbb{E}[aX] &= \sum_x axP_X(x) \\
&= a\sum_x xP_X(x) \\
&= a\mathbb{E}[X]
\end{aligned}
$$

So, we can assume $a = b = 1$ WLOG.

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

This holds without *any* assumption of independence, etc. □

Example 3.8

Let X_1, X_2 denote outcomes of rolling two fair dice.

Linearity of expectation gives

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

Example 3.9

Let X_1 be a roll of a fair dice, and X_2 equal to $7 -$ (first roll).

Linearity of expectation gives (similarly)

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

A really powerful technique is to introduce "indicator" random variables and use linearity of expectation.

A famous example is the hat example:

Example 3.10

n people put their hats into a basket, and draw a random one out. What is the expected number of people who get their own hat back?

Let the indicator RVs

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

$$\begin{aligned} \mathbb{E}\big[\text{number of people that get their own hat}\big] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}[X_i] \end{aligned}$$

where

$$\mathbb{E}[X_i] = 1 \cdot \mathbb{P}\big(\text{person } i \text{ gets their own hat}\big) + 0 \cdot \mathbb{P}\big(\text{person } i \text{ gets another hat}\big) = \frac{1}{n}.$$

This means that the final expected value is just

$$\sum_{i=1}^{n} \mathbb{E}[X_i] = n \cdot \frac{1}{n} = 1.$$

Introducing random variables bypasses the need for combinatorics, and simplifies the problem.

3.4 Tail Sum Formula for Expectation

Theorem 3.11: Tail Sum for Expectation

For nonnegative integer-valued (integers for simplicity) random variables, we have

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

Proof.

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

□

Example 3.12

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

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

1/28/2021

Lecture 4

Variance, Discrete Distributions, Conditional Probability

Definition 4.1: Variance

The variance of a random variable X describes the "spread" of X around its expectation.

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

A useful fact is that if X, Y are independent (i.e. $P_{XY}(x,y) = P_X(x)P_Y(y)$), then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. We'd need to prove the following lemma first though:

Lemma 4.2

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

Proof.

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

□

With this lemma, we can prove the earlier fact:

Lemma 4.3: Sum of Variances

If X, Y are independent (i.e. $P_{XY}(x,y) = P_X(x)P_Y(y)$), then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof.

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

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

□

Definition 4.4: Covariance

The covariance between two RVs X and Y is defined as

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

It's a useful notion of dependence between X and Y; $\mathrm{Cov}(X, Y) = 0 \iff X, Y$ are uncorrelated.

Note: uncorrelated is not the same as independent, but independence implies not correlated.

Definition 4.5: Correlation

The correlation coefficient is

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

This coefficient is always between -1 and 1.

Why is it always between -1 and 1? It's Cauchy-Schwarz in disguise:

$$
\begin{aligned}
\mathbb{E}[XY] &= \sum_{x,y} xy P_{XY}(x, y) \\
&= \sum_{x,y} x P_{XY}(x, y)^{\frac{1}{2}} y P_{XY}(x, y)^{\frac{1}{2}} \\
&\leq \left(\sum_{x,y} x^2 P_{XY}(x, y)\right)^{\frac{1}{2}} \left(\sum_{x,y} y^2 P_{XY}(x, y)\right)^{\frac{1}{2}} \\
&= \left(\sum_{x} x^2 P_X(x)\right)^{\frac{1}{2}} \left(\sum_{y} y^2 P_Y(y)\right)^{\frac{1}{2}} \\
&\implies \left|\rho(X, Y)\right| \leq 1
\end{aligned}
$$

4.1 Common Discrete Distributions

- $X \sim \mathrm{Uniform}(\{1, 2, \ldots, n\}) \implies P_X(k) = \begin{cases} \frac{1}{n} & k = 1, \ldots n \\ 0 & \text{otherwise} \end{cases}$
- $X \sim \mathrm{Bernoulli}(p) \implies P_X(k) = \begin{cases} 1 - p & k = 0 \\ p & k = 1 \\ 0 & \text{otherwise} \end{cases}$

  $\mathbb{E}[X] = 0(1 - p) + 1(p) = p.$

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

  Note that $X = \sum_{i=1}^{n} X_i$, where $X_i \sim_{IID} \mathrm{Bernoulli}(p)$.

Using this, we can see that $\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^n X_i\right] = np$ by linearity of expectation, and $\text{Var}(X) = \text{Var}(\sum_{i=1}^n X_i) = np(1-p)$ because X_i are independent.

- Indicator random variables of $A \in \mathcal{F}$: $\mathbf{1}_A \sim \text{Bernoulli}(P(A))$
- $X \sim \text{Geom}(p) \implies P_X(k) = \begin{cases} p(1-p)^{k-1} & k = 1,2,\dots \\ 0 & \text{otherwise} \end{cases}$

  X = number of trials (each independent Bernoulli(p) until we get a success).

  Equivalently, $X = \min\{k \geq 1 \mid X_k = 1\}$, $X_i \sim_{IID} \text{Bernoulli}(p)$.

  $\mathbb{E}[X] = \sum_{k\geq 1} \mathbb{P}(X \geq k) = \sum_{k\geq 1}(1-p)^{k-1} = \frac{1}{p}$ (as it's a geometric series).

  Similarly, $\text{Var}(X) = \frac{1-p}{p^2}$.
- $X \sim \text{Pois}(\lambda)$, where λ is the "rate" parameter > 0.

  $P_X(k) = \begin{cases} \frac{\lambda^k e^{-\lambda}}{k!} & k = 0,1,2,\dots \\ 0 & \text{otherwise} \end{cases}$, and $\mathbb{E}[X] = \lambda$.

  Useful in modeling the number of "arrivals". Where does Poisson come from?

  Let $X_n \sim \text{Bin}(n, p_n)$, where $\mathbb{E}[X_n] = np_n \to \lambda$ as $n \to \infty$.

  Our claim is that $\mathbb{P}(X_n = k) \to \frac{\lambda^k e^{-\lambda}}{k!}$ (the Poisson law of rare events).

  The intuition behind this is to imagine n time intervals in an hour. An arrival in interval i is distributed as Bernoulli(p_n), and the total number of arrivals $\sim \text{Bin}(n, p_n)$.

  As $n \to \infty$, we have

$$\begin{aligned} \mathbb{P}(X_n = k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} \\ &\to \frac{\lambda^k}{k!} e^{-\lambda} \end{aligned}$$

  The last equality makes use of $(1-\frac{\lambda}{n})^n \approx e^{-\lambda}$ as $n \to \infty$, and also $n(n-1)\cdots(n-k+1) \approx n^k$ as $n \to \infty$.

Note that by construction of Binomial distributions, if $X \sim \text{Bin}(n,p)$ and $Y \sim \text{Bin}(m,p)$ are independent, then $X + Y \sim \text{Bin}(m+n, p)$.

Similarly, if $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\mu)$ are independent, then $X + Y \sim \text{Pois}(\lambda + \mu)$. An interesting consequence of this is that different classes of arrivals can be combined.

Example 4.6: Coupon Collector Problem

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

Let X_i = the number of boxes I need to buy to get the ith new coupon, starting from when I find the $(i-1)$th coupon.

We have that $X_i \sim \text{Geom}(\frac{N-i+1}{N})$, because we've already seen $i-1$ coupons, so we have $N-i+1$ left.

This means that by linearity of expectation, $\mathbb{E}[\#\text{ boxes to buy}] = \sum_{i=1}^N \mathbb{E}[X_i] = \frac{N}{N} + \frac{N}{N-1} + \cdots + \frac{N}{1} \approx N \log N$ for N large.

4.2 Conditional Probability

Recall that $\mathbb{P}(A \mid C) = \frac{\mathbb{P}(A \cap C)}{P(C)}$, where $P(C) > 0$ (from Definition 2.3).

For discrete RVs, we can define the conditional distribution (conditional pmf)

$$P_{X|Y}(x \mid y) = \frac{P_{XY}(x,y)}{P_Y(y)} = \frac{\mathbb{P}\big(X = x \cap Y = y\big)}{\mathbb{P}\big(Y = y\big)} = \mathbb{P}\big(X = x \mid Y = y\big).$$

This makes sense, provided $P_Y(y) > 0$. Note that this also means that for each fixed y, $P_{X|Y}(\cdot \mid y)$ is a pmf on X.

The interpretation of this would be: given $Y = y$, what is the new distribution over X?

Example 4.7

We pick up coin 1 with probability $\frac{1}{2}$ (bias is $\frac{1}{4}$), and we pick up coin 2 with probability $\frac{1}{2}$ (bias is $\frac{3}{4}$). We toss the coin twice (suppose $H = 1$ and $T = 0$). Let Y = first toss, and X = second toss.

$$\begin{aligned} P_{X|Y}(1 \mid 1) &= \frac{P_{XY}(1,1)}{P_Y(1)} \\ &= \frac{\frac{1}{2} \cdot \left(\frac{1}{4}\right)^2 + \frac{1}{2}\left(\frac{3}{4}\right)^2}{\frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{3}{4}} \\ &= \frac{5}{8} \end{aligned}$$

The geometric RV has a memoryless property; if $X \sim \text{Geom}(p)$, then

$$\mathbb{P}(X = k + m \mid X > k) = \mathbb{P}(X = m).$$

This makes sense, because if I've already tried k times, at this point it's just like if I just showed up and began.

2/2/2021

Lecture 5

Conditional Expectation, Continuous Random Variables

5.1 Conditional Expectation

Definition 5.1: Conditional Expectation

The conditional expectation is defined as

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

This also represents the expected value of X given that a I know $Y = y$.

Note that $\mathbb{E}\big[X \mid Y = y\big]$ is a function of y.

As we saw before, if Z is a RV, so is $g(Z)$. So if we were to evaluate the expectation at a *random value* of y, we have $\mathbb{E}[X \mid Y] = g(Y)$, which is a random variable (a function of Y).

The most important property of conditional expectation:

Theorem 5.2: Tower Property

For all functions f,

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

Proof.

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

□

This property also allows us to "iterate" the expectations.

Theorem 5.3: Iterated Expectation

Iterated expectation follows directly from taking $f(Y) = 1$ in the tower property.

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

Example 5.4

Toss a fair coin N times, and let H = number of heads; N and H are RVs.

$$\mathbb{E}[H] = \mathbb{E}[\mathbb{E}[H \mid N]] = \mathbb{E}\left[\frac{N}{2}\right] = \frac{1}{2}\mathbb{E}[N].$$

5.2 Continuous Random Variables

$X \sim \text{Uniform}(0,1) \implies X$ is a random number between 0 and 1.

Recall for discrete RVs, the pmf defines its distribution, i.e.

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

For a continuous RV, its distribution is defined by its "density" $f_X : \mathbb{R} \to [0,\infty)$:

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

Densities must satisfy: $f_X \geq 0$ and $\int_{\mathbb{R}} f_X(x)\,\mathrm{d}x = 1$. This is to satisfy the axioms of probability.

Observe that for continuous RVs, $\mathbb{P}(X = x) = 0$.

Why?

$$\mathbb{P}(X = x) \leq \mathbb{P}(x \leq X < x + \delta) = \int_x^{x+\delta} f_X(u)\,\mathrm{d}u \approx \delta f_X(x).$$

As $\delta \to 0$, $\mathbb{P}(X = x) = 0$.

Definition 5.5: Cumulative Distribution Function (cdf)

The cumulative distribution function (cdf) of a random variable X is defined as

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

For a continuous RV,

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

Properties of cdfs in general:

- $\lim_{x\to-\infty} F_X(x) = 0$, because $\mathbb{P}(X \le -\infty) = 0$
- $\lim_{x\to\infty} F_X(x) = 1$, because $\mathbb{P}(X \le \infty) = 1$
- $\lim_{y\downarrow x} F_X(y) = F_X(x)$; $y \downarrow x$ means that y is descending to x—this is also called right-continuity.

5.3 Examples of Continuous RVs

- $X \sim \text{Uniform}(a,b) \implies f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$

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

- $X \sim \text{Exp}(\lambda) \implies f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & \text{otherwise} \end{cases}$

  $F_X(x) = \lambda \int_0^x e^{-\lambda u}\,\mathrm{d}u = 1 - e^{-\lambda x}$

  The exponential distribution is the continuous analog of a geometric distribution; it also has the memoryless property.

5.4 Extensions to multiple RVs

We say $X_1, X_2, \ldots, X_n$ are (jointly) continuous RVs if there is a function $f_{X_1X_2\cdots X_n} : \mathbb{R}^n \to [0,\infty)$ such that

$$\begin{aligned} F_{X_1\cdots X_n}(x_1,x_2,\ldots,x_n) &= \mathbb{P}(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n) \\ &= \int_{-\infty}^{x_n}\int_{-\infty}^{x_{n-1}} \cdots \int_{-\infty}^{x_1} f_{X_1\cdots X_n}(u_1,\ldots,u_n)\,\mathrm{d}u_1\cdots\mathrm{d}u_{n-1}\,\mathrm{d}u_n \end{aligned}$$

Example 5.6

If we throw a dart at a dartboard of radius r, let (X,Y) denote the pair of xy coordinates of where the dart lands.

If the dart lands uniformly on the board, what is f_{XY}?

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

RVs X, Y are independent if

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

or in other words, the events $\{X \le x\}$ and $\{Y \le y\}$ are independent events.

If X, Y are continuous, this is equivalent to

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

In the dartboard example, X, Y are not independent; knowing X limits the values of Y.

If instead (X,Y) were uniform on the box $[0,r] \times [0,r]$, then X, Y are independent:

$$f_{XY}(x,y) = \frac{1}{r^2} \mathbf{1}_{\{x \in [0,r] \cap y \in [0,r]\}} = f_X(x) f_Y(y).$$

For continuous RVs, expectation, etc. are similar to the discrete case; we just replace sums by integrals.

For example, $\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x$. More generally,

$$\mathbb{E}\big[g(X_1,\dots,X_n)\big] = \int \cdots \int_{\mathbb{R}^n} g(x_1,\dots,x_n) f_{X_1,\cdots X_n}(x_1,\cdots x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

For variance,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] \\
&= \mathbb{E}\big[X^2\big] - \mathbb{E}[X]^2 \\
&= \int x^2 f_X(x)\,\mathrm{d}x - \left(\int x f_X(x)\,\mathrm{d}x\right)^2
\end{aligned}$$

Example 5.7

If $X \sim \mathrm{Uniform}(a,b)$, then

$$\mathbb{E}[X] = \frac{1}{b-a}\int_a^b x\,\mathrm{d}x = \frac{1}{2}\frac{b^2-a^2}{b-a} = \frac{1}{2}(b+a).$$

Doing similar computations, we also have $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$

Example 5.8

Going back to the dartboard example, let $R = \sqrt{X^2+Y^2}$ = distance from the dart to the center of the board.

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

5.5 Conditioning for continuous RVs

Let X, Y be continuous RVs. The conditional density of X given $Y = y$ is defined as

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

Note that this is a device to simplify calculations; it's not a true "conditional probability", in a sense that $\mathbb{P}\big(Y = y\big) = 0$, though it has positive density. Its existence makes sense intuitively, but there is some deeper stuff going on; it's a special case of "disintegration of measure".

With conditional density, we can define conditional expectation for continuous RVs.

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

Similarly, $\mathbb{E}[X \mid Y]$ denotes this function evaluated at Y.

Note that the tower property still holds here.

Example 5.9

Back to the dartboard example, we have $\mathbb{E}[X \mid Y] = 0$. This makes sense, because the X position is still symmetrical around 0 even for any given Y value.

2/4/2021

Lecture 6

Memoryless Property, Gaussian Random Variables

6.1 Memoryless Property

Suppose we want a continuous RV X with a "memoryless" property like that of the geometric RV.

Mathematically, we want a RV X such that

$$\mathbb{P}(X > t + s \mid X > s) = \mathbb{P}(X > t),$$

for $s, t \geq 0$.

This means that

$$\begin{aligned}\frac{\mathbb{P}(X > t + s \cap X > s)}{\mathbb{P}(X > s)} &= \mathbb{P}(X > t)\\ \frac{\mathbb{P}(X > t + s)}{\mathbb{P}(X > s)} &= \mathbb{P}(X > t)\\ \mathbb{P}(X > t + s) &= \mathbb{P}(X > t)\,\mathbb{P}(X > s)\end{aligned}$$

This means we need a function $f(t+s) = f(t) \cdot f(s)$; the only nonzero solution to this is $f(t) = e^{\alpha t}$ for some α.

This means that the cdf $F_X(t) = 1 - \mathbb{P}(X > t) = 1 - e^{\alpha t}$ for some $\alpha < 0$. This means that we can conclude that the only RVs with this property have a cdf of the form

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

This is the cdf of $\mathrm{Exp}(\lambda)$.

6.2 Gaussian Random Variables

Like exponential RVs, Gaussians emerge naturally in many ways, and are another important class of RVs.

The central limit theorem (which we'll talk about later) says that the sum of many small independent effects is gaussian.

Definition 6.1: Gaussian Random Variable

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

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

For $X \sim \mathcal{N}(0,1)$ (standard normal), we define the cdf as

$$\Phi(x) = F_{\mathcal{N}(0,1)}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u.$$

There is no closed form for Φ.

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

Is f_X actually a density? It's nonnegative, but we need to check if it integrates to 1.

Let's assume WLOG that $\mu = 0$ and $\sigma^2 = 1$.

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

Some cool facts about gaussians:

- If $X \sim \mathcal{N}(\mu_X \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are independent, then $X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

  Caution: independence is needed here in general.
- If X, Y are independent and $X + Y \sim \mathcal{N}(\mu, \sigma^2)$, then X and Y are both gaussian.
- If X, Y are independent and $X + Y$ and $X - Y$ are independent, then both X and Y are gaussian with the same variance.

Example 6.2

T = temperature of a chip on a satellite, averaged over 1 second, and let $T \sim \mathcal{N}(20, 2)$. There is a failure if the temperature goes below 10° or above 50° in a 1 second interval.

$$\begin{aligned}
\mathbb{P}(\text{failure}) &= \mathbb{P}(T < 10) + \mathbb{P}(T > 50) \\
&= \mathbb{P}\left(\frac{T-20}{\sqrt{2}} < \frac{10-20}{\sqrt{2}}\right) + \mathbb{P}\left(\frac{T-20}{\sqrt{2}} > \frac{50-20}{\sqrt{2}}\right) \\
&= \Phi\left(-\frac{10}{\sqrt{2}}\right) + 1 - \Phi\left(\frac{30}{\sqrt{2}}\right) \\
&\approx 7.7 \times 10^{-13} + 3.6 \times 10^{-100} \\
&\approx 7.7 \times 10^{-13}
\end{aligned}$$

What's the probability that we fail after 25 years of spaceflight?

There are 7.9×10^8 seconds in 25 years, so by union bound,

$$\begin{aligned}\mathbb{P}(\text{fail after 25 years}) &\leq 7.9 \times 10^8 \cdot 7.7 \times 10^{-8} \\ &< \frac{1}{1000}\end{aligned}$$

6.3 Mixing Discrete and Continuous RVs

We've seen (X, Y) discrete and (X, Y) continuous; what about X discrete and Y continuous?

The key idea is to factor into conditional and marginal distributions.

Example 6.3

A cellphone sends a bit $B \in \{-1, +1\}$ (equally likely) to the tower. The tower receives $Y = B + N$, where $N \sim \mathcal{N}(0,1)$, independent of B. (Here, N can model noise, interference, multipath, etc.)

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

where $y \in \mathbb{R}$ and $b \in \{-1, +1\}$.

Given that I observe $Y = y$, what is the probability that $B = +1$ was sent?

By Bayes rule,

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

2/9/2021

Lecture 7

Conditional Variance, Derived Distributions

7.1 Conditional Variance

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

Conditional variance is basically the same. That is, $\text{Var}(X \mid Y = y)$ is the variance of X given $Y = y$, or

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

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

Note that $\mathbb{E}[\text{Var}(X \mid Y)] = \mathbb{E}\left[(X - \mathbb{E}[X \mid Y])^2\right]$ by iterated expectation. The latter is called the minimum-mean-square error (MMSE) of X given Y.

Theorem 7.1: Law of total variance

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

Proof.

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\
&= \mathbb{E}\left[\mathbb{E}\left[X^2 \mid Y\right]\right] - \mathbb{E}[\mathbb{E}[X \mid Y]]^2 \\
&= \mathbb{E}\left[\text{Var}(X \mid Y) + \mathbb{E}[X \mid Y]^2\right] - \mathbb{E}[\mathbb{E}[X \mid Y]]^2 && \text{(by definition of conditional variance)} \\
&= \mathbb{E}[\text{Var}(X \mid Y)] + \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right] - \mathbb{E}[\mathbb{E}[X \mid Y]]^2 \\
&= \mathbb{E}[\text{Var}(X \mid Y)] + \text{Var}(\mathbb{E}[X \mid Y]) && \text{(by definition of variance)}
\end{aligned}$$

□

Note that this proof works entirely by multiplying things at the level of variance and expectation. We never needed to work at the level of individual pmfs/pdfs, etc.

Remark: the MMSE of X given $Y = \text{Var}(X)$ is equal to $\text{Var}(\mathbb{E}[X \mid Y])$ (rearranging the law of total variance). The term $\text{Var}(\mathbb{E}[X \mid Y])$ can be interpreted as how much "uncertainty" about X is reduced by knowing Y (on average).

Example 7.2

Let $Y = X_1 + X_2 + \cdot + X_N$, where X_i's are IID, and N is a RV taking values in $\{1, 2, \ldots\}$. What is $\text{Var}(Y)$?

By the law of total variance, we have

$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y \mid N)] + \text{Var}(\mathbb{E}[Y \mid N]).$$

If we're given N, then $\text{Var}(Y) = N\text{Var}(X_i)$ because the X_i's are independent. Similarly, $\mathbb{E}[Y \mid N] = N\mathbb{E}[X_i]$ because of linearity of expectation. This gives us

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

7.2 Derived Distributions

The distribution of a RV X is described by its cdf F_X. What if we have a new RV $Y = g(X)$ for some function g? How do we find the distribution of Y? The distribution of Y is called a *derived distribution*, because its distribution is derived from X.

First, ask yourself: do I really need the distribution of Y? Oftentimes, you don't. For example, $\mathbb{E}[Y] = \mathbb{E}\left[g(X)\right]$, which is just a function of the distribution of X, and LOTUS applies. More generally, $\mathbb{E}\left[f(Y)\right] = \mathbb{E}\left[(f \circ g)(X)\right]$.

If you really do want the distribution of Y, it's best to work with cdfs.

$$\begin{aligned}
F_Y(y) &= \mathbb{P}\left(Y \le y\right) \\
&= \mathbb{P}\left(g(X) \le y\right) \\
&= \mathbb{P}\left(X \in g^{-1}(-\infty, y]\right) = \mathbb{P}\left(X \in \{x \mid g(x) \le y\}\right)
\end{aligned}$$

This last term has no closed form in general.

In the case of discrete random variables, we have

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

If g is invertible, then $P_Y(y) = P_X(g^{-1}(y))$. Note that this formula only works for discrete distributions. For continuous distributions, it's a bit more complicated.

Example 7.3

Let $f_X(x) = \begin{cases} 1 & 0 \le x < 1 \\ 0 & \text{otherwise} \end{cases}$, and $Y = 2X$.

Blindly applying the discrete formula suggests that $f_Y(y) = f_X\left(\frac{y}{2}\right) = \begin{cases} 1 & 0 \le y < 2 \\ 0 & \text{otherwise} \end{cases}$.

We have a problem here—the thing on the right doesn't integrate to 1, so it's not a valid density function.

It's better to work at the level of cdfs:

$$\begin{aligned} F_Y(y) &= \mathbb{P}\big(Y \le y\big) \\ &= \mathbb{P}\big(2X \le y\big) = \mathbb{P}\Big(X \le \frac{y}{2}\Big) \\ &= F_X\Big(\frac{y}{2}\Big) \end{aligned}$$

We can then compute $f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X\left(\frac{y}{2}\right) = \frac{1}{2} f_X\left(\frac{y}{2}\right) = \begin{cases} \frac{1}{2} & 0 \le y < 2 \\ 0 & \text{otherwise} \end{cases}$.

Example 7.4: Order Statistics

Let $X_1, \dots, X_n$ be IID, and sort them so that $X^{(1)} \le X^{(2)} \le \cdots \le X^{(n)}$, such that $X^{(i)}$ is the ith smallest number in the list $(X_1, X_2, \dots, X_n)$. If X_i are continuous random variables with density f_X, what is the density of $X^{(i)}$?

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

"Proof". We have

$$f_{X^{(i)}}(y)\,\mathrm{d}y \approx \mathbb{P}\Big(X^{(i)} \in (y, y + \mathrm{d}y)\Big),$$

and explaining each part of the formula, we have

$$\overbrace{n}^{\substack{\text{ways to select } X_i \\ \text{that fall in} \\ [y, y+\mathrm{d}y]}} \quad \underbrace{\binom{n-1}{i-1}}_{\substack{\text{ways to choose } i-1 \\ \text{of remaining } n-1 \\ X_i\text{s to be} \le y}} \quad \overbrace{F_X(y)^{i-1}}^{\substack{\text{probability } i-1 \\ X_i\text{s are} \le y}} \quad \underbrace{\big(1 - F_X(y)\big)^{n-i}}_{\substack{\text{probability } n-i \\ X_i\text{s are} \ge y+\mathrm{d}y}} \quad \overbrace{f_X(y)\,\mathrm{d}y}^{\substack{\text{probability one of} \\ \text{the } X_i \in [y, y+\mathrm{d}y]}} .$$

□

Example 7.5

Suppose I model the time to bus of type $k = 1, \dots, n$ arriving as $X_k \sim \mathrm{Exp}(\lambda)$, all independent.

When does the ith soonest bus arrive? At time $X^{(i)}$. The distribution in this case is

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

Another common example of a derived distribution is the sum of two independent RVs.

Example 7.6: Convolutions

If X, Y are discrete, integer-valued independent RVs, and $Z = X + Y$, what is the pmf of Z?

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

This is called the *convolution* $(P_X * P_Y)(z)$.

Similarly, if X, Y are independent continuous RVs with densities f_X, f_Y, then $Z = X + Y$ has density

$$f_Z(z) = \int f_X(x) f_Y(z - x)\,\mathrm{d}x = \big(f_X * f_Y\big)(z).$$

The takeaway here is that adding independent RVs corresponds to the convolution of distributions.

2/11/2021

Lecture 8

Moment Generating Functions, Concentration Inequalities

As an aside:

Theorem 8.1: Independence of functions of RVs

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

Proof. By definition, if X and Y are independent, then the events $\{X \in A\}$ and $\{Y \in B\}$ are independent. That is,

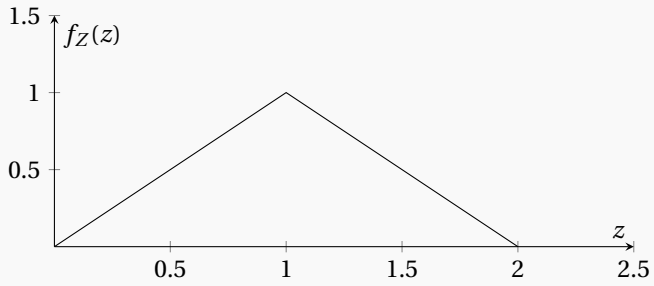
$$\mathbb{P}(X \in A \cap Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B).$$

If we have $f(X)$ and $f(Y)$, we have to show that $\{f(X) \in E\}$ and $\{f(Y) \in F\}$ are independent. We can express these events equivalently as $\{X \in f^{-1}(E)\}$ and $\{Y \in f^{-1}(F)\}$, respectively. We can just define $A = f^{-1}(E)$ and $B = f^{-1}(F)$, and we've shown that these two events are also independent.

Therefore, $f(X)$ and $f(Y)$ are also independent. □

Example 8.2

If $X, Y \sim \text{Uniform}(0, 1)$, then we get this distribution of $Z = X + Y$:



In general, computing convolutions can be tedious. In 120 (for example), you will have seen that computations are simplified by working in an appropriate "transform domain".

This motivates the definition of "moment generating functions".

8.1 Moment Generating Functions

Definition 8.3: Moment Generating Function (MGF)

For $t \in \mathbb{R}$, provided that the expectation exists, the moment generating function of X is defined as

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

Why "moment generating function"?

A generating function takes a sequence of numbers $(a_0, a_1, a_2, \cdots)$ and transforms it into some polynomial/power series $\sum_{k\geq 0} a_k z^k$; convolutions of sequences correspond to multiplication of polynomials.

If we expand the earlier definition of moment generating functions, we have

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

Since $\mathbb{E}[X^n]$ is the nth moment of X, MGFs encode moments of a distribution in coefficients of a power series.

An important fact: If the MGF exists, then it uniquely determines the distribution of X.

To recover moments from the MGF, we have, by the power series above,

$$\frac{\mathrm{d}}{\mathrm{d}t} M_X(t)\bigg|_{t=0} = \mathbb{E}[X].$$

Similarly,

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n} M_X(t)\bigg|_{t=0} = \mathbb{E}\left[X^n\right].$$

Example 8.4

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

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

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

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

This is precisely the MGF of $\mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

This makes calculating sums of RVs a lot easier; we just need to multiply the MGFs instead of convolving the pmfs.

But in principle, everything you can do with MGFs, you can also do by working directly with the distributions.

As an aside (out of scope): More generally, one usually uses characteristic functions instead of MGFs:

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

for $t \in \mathbb{R}$. $\phi_X(t)$ is the "characteristic function" of X. The RHS is the Fourier transform of the distribution of X. This always exists, and uniquely characterizes the distribution.

8.1.1 Examples of MGFs

- $X \sim \mathcal{N}(\mu, \sigma^2)$: $M_X(t) = \exp\left(t\mu + t^2\frac{\sigma^2}{2}\right)$
- $X \sim \text{Exp}(\lambda)$: $M_X(t) = \frac{\lambda}{\lambda - t}$ for $t < \lambda$
- $X \sim \text{Pois}(\lambda)$: $M_X(t) = e^{-\lambda + \lambda e^t}$
- $X \sim \text{Geom}(p)$: $M_X(t) = \frac{p^t}{1-(1-p)e^t}$ for $t < \ln(1-p)$
- $X \sim \text{Bin}(n, p)$: $M_X(t) = \left(1 - p + pe^t\right)^n$
- $X \sim \text{Bernoulli}(p)$: $M_X(t) = (1-p) + pe^t$

Example 8.5

Let $X_i \sim_{IID} \text{Bernoulli}(p)$. We know that $Y = \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$ by the definition of binomial distributions.

We can verify that this is the case by using MGFs:

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^n M_{X_i}(t) \\ &= \prod_{i=1}^n \left(1 - p + pe^t\right) \\ &= \left(1 - p + pe^t\right)^n \end{aligned}$$

This is precisely the MGF of $\text{Bin}(n, p)$, so $Y \sim \text{Bin}(n, p)$.

8.2 Concentration Inequalities

Seldom in applications can we compute probabilities of interesting events in closed form. It's also usually not necessary to compute them like this. Usually, we settle for an inequality.

In particular, we usually want to show $P(A) \approx 0$ or $P(A) \approx 1$.

In a lot of randomness, there is (almost) determinism; the whole point of probability is arguably to (almost) deterministically predict an outcome. Concentration inequalities are just one step in doing this.

Example 8.6

Let $X_i \sim_{IID} X$, and define the empirical mean

$$M_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

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

Since the variance is vanishing, M_n must be "almost constant" for $n \gg 1$.

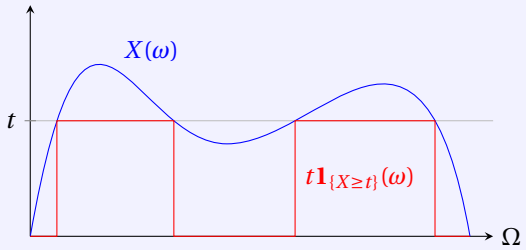
Can we say more?

Theorem 8.7: Markov Inequality

If X is a nonnegative random variable, then for $t > 0$,

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

Proof. We claim that $X \geq t\mathbf{1}_{\{X \geq t\}}$, or as a function of ω, $X(\omega) \geq t\mathbf{1}_{\{X \geq t\}}(\omega)$.



This means that we have

$$\mathbb{E}[X] \geq t\mathbb{E}\left[\mathbf{1}_{\{X \geq t\}}\right] = t\mathbb{P}(X \geq t)$$

□

It's important to remember/understand Markov's Inequality because the rest are essentially just consequences of Markov's Inequality.

Theorem 8.8: Chebyshev's Inequality

If X is a random variable, then

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

Proof.

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

□

Theorem 8.9: Weak Law of Large Numbers

Let $X_1, X_2, \cdots \sim_{IID} X$, and define $M_n = \frac{X_1 + \cdots + X_n}{n}$.

For $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}(|M_n - \mathbb{E}[X]| > \varepsilon) = 0.$$

Proof. We know that $\mathbb{E}[M_n] = \mathbb{E}[X]$ by linearity of expectation, and $\text{Var}(M_n) = \frac{\text{Var}(X)}{n}$.

By Chebyshev, we have $\mathbb{P}(|M_n - \mathbb{E}[X]| > \varepsilon) \leq \frac{\text{Var}(X)}{n\varepsilon^2}$, which tends to 0 as $n \to \infty$. □

Let $X_1, X_2, X_3 \cdot \sim_{IID} X$. Think about these X_i as outcomes of an experiment modeled by X, repeated under identical conditions.

We can define the empirical frequency of $\{X_i \in B\}$ as

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

By WLLN, we have as $n \to \infty$,

$$\mathbb{P}(|F_n - \mathbb{P}(X \in B)| > \varepsilon) \to 0.$$

In other words, $\mathbb{P}(X \in B)$ *is* the frequency at which X takes values in B under repeated trials. This is something we've been taking as fact, but we've just proven it as well.

This proof means that our axiomatic framework is compatible with the "frequentist" or "empirical" framework.

2/16/2021

Lecture 9

Chernoff Bound, Convergence

9.1 Chernoff Bounds

Theorem 9.1: Chernoff Bound

For a RV X, $a \in \mathbb{R}$, and $t > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}\left[e^{tx}\right]}{e^{ta}} = e^{-ta} M_X(t).$$

Since the RHS is true for all $t > 0$, we can optimize over $t > 0$ to get the best bound.

Proof.

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

□

In a sense, Chebyshev gives a "better" bound than Markov because it uses the additional info of the second moment (for $\text{Var}(X)$). Chernoff (with the MGF), is using information about *all* moments of X.

Example 9.2

Let $Z \sim \mathcal{N}(0,1)$. We know that the cdf $\Phi(x)$ has no closed form expression, but Chernoff gives us good control of tail probabilities.

For $a > 0$, Chernoff gives

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

If we optimize this (i.e. $t = a$ is the best choice as it's the minimum), we have

$$\mathbb{P}(Z \geq a) \leq e^{-a^2/2}.$$

9.2 Convergence of Random Variables

Convergence is essentially the language of "limits" in probability.

For a sequence of real numbers $a_1, a_2, \cdots \in \mathbb{R}$, we know what it means to write $\lim_{n\to\infty} a_n = a$.

Given a sequence of RVs $X_1, X_2, \cdots$, what does it mean to write $\lim_{n\to\infty} X_n = X$? Nothing.

Why? Random variables are *functions*. So, for a sequence of functions $f_1, f_2, \cdots$, what does it mean to write $\lim_{n\to\infty} f_n = f$? We need more information, because there are many ways to talk about convergence of functions.

For example,

- Pointwise: $\lim_{n\to\infty} f_n(x) = f(x)$, $\forall x$
- In L^1-norm: $\lim_{n\to\infty} \int \left| f_n(x) - f(x) \right| \mathrm{d}x = 0$

9.2.1 Modes of convergence

There are three useful "modes" of convergence:

Definition 9.3: Almost Sure Convergence

We say that $X_n \to X$ almost surely if $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$.

This is also notated as $X_n \xrightarrow{a.s.} X$.

The LHS is equivalent to $\mathbb{P}(\{\omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\})$, i.e. as functions X_n converge pointwise to X on some set of samples A having $P(A) = 1$.

Definition 9.4: Convergence in Probability

We say that $X_n \to X$ in probability if for every $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0.$$

This is also notated as $X_n \xrightarrow{p} X$.

For example, WLLN tells us that the empirical mean $\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}[X]$ in probability.

Definition 9.5: Convergence in Distriubtion

We say that $X_n \to X$ in distribution if the cdfs converge, i.e. for all continuity points x of F_X,

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

This is also notated as $X_n \xrightarrow{d} X$.

Relationship among modes of convergence:

$$(X_n \to X \text{ a.s.}) \implies \big(X_n \to X \text{ in prob.}\big) \implies (X_n \to X \text{ in dist.}).$$

All of these implications are strict.

Example 9.6

Let $X_1, X_2, \cdots \sim_{IID} \text{Bernoulli}(\frac{1}{2})$.

$X_n \to X \sim \text{Bernoulli}(\frac{1}{2})$ in distribution (trivially), but $X_n \not\to$ anything a.s.

Why? Because in order to converge to something, we need a sequence with finitely many ones or zeros, which occurs with zero probability.

Theorem 9.7: Strong Law of Large Numbers

If $X_1, X_2, \cdots \sim_{IID} X$ and $\mathbb{E}[X] < \infty$, then

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

This is the same as WLLN, but claims a.s. convergence instead of convergence in probability, i.e. SLLN implies WLLN.

SLLN tells us that individual sample paths $\frac{1}{n}\sum_{i=1}^{n} X_n(\omega) \to \mathbb{E}[X]$ for all ω in some event having probability 1.

SLLN doesn't tell us anomalies *can't* happen, but it tells us they *won't* happen. For example, if we have a sequence of coin flips, it's *possible* that we have a sequence of all heads, but SLLN says that we will never observe this.

Theorem 9.8: Central Limit Theorem

Let $X_1, X_2, \cdots \sim_{IID} X$, with $\text{Var}(X) = \sigma^2 < \infty$, and $\mathbb{E}[X] = \mu$.

Define $S_n = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n}\sigma}$. Then, $S_n \to Z \sim \mathcal{N}(0,1)$ in distribution.

In other words, $\mathbb{P}(S_n \le x) \to \Phi(x)$, $\forall x \in \mathbb{R}$.

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

$$M_X(t) = \sum_{n \ge 0} t^n \frac{\mathbb{E}[X^n]}{n!} = 1 + \frac{t^2}{2} + o\left(t^2\right).$$

Where $o(t^2)$ means that it's vanishing faster than t^2 as $t^2 \to 0$.

We also have

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

Where the last equality comes from $\lim_{n\to\infty}\left(1 + \frac{c}{n}\right)^n = e^c$. This last expression is the exact same as the mgf of a standard normal, i.e.

$$\lim_{n\to\infty} M_{S_n}(t) = M_{\mathcal{N}(0,1)}(t).$$

This implies that $\lim_{n\to\infty} S_n = \mathcal{N}(0,1)$ in distribution. This pointwise convergence of characteristic function implies convergence in distribution is nontrivial (it's known as Lévy's continuity theorem). □

A word of caution: $X_n \to X$ a.s. does *not* imply that $\mathbb{E}[X_n] \to \mathbb{E}[X]$. We need more information to conclude this.

2/23/2021

Lecture 10

Information Theory, Source Coding

Now that we have covered the basics of probability frameworks, we'll put them to good use.

10.1 Information Theory

The entire field of information theory can be traced back to one paper: *Mathematical Theory of Communication* by Shannon in 1948.

This singular paper kicked off the "information age".

Some questions addressed in the paper:

1. How reliably and how quickly can I communicate a message over a noisy channel?

   Example 10.1: Cocktail Party

   Suppose you're at a party, and it's really loud, so what you say is being corrupted by the noise in the room. What you can do is either speak louder, speak slower, repeat what I'm saying, etc.

   This is also called the channel coding problem.

2. How many bits do I need to losslessly represent an observation?

   Example 10.2: Data Compression

   If we have an image, there's a lot of data redundancy, and these redundancies are compressed, ex. in the JPEG standard.

   This is also called the source coding problem.

Shannon's mathematical insights guided decades of development in these areas.

10.2 Source Coding (Compression)

For a discrete random variable X with pmf P_X, we define the (Shannon) entropy as

$$H(X) = \sum_x P_X(x) \log\left(\frac{1}{P_X(x)}\right) = \mathbb{E}\left[\log\left(\frac{1}{P_X(X)}\right)\right].$$

We typically take the base of the logarithm to be 2, so the units of entropy are "bits".

What is entropy? We can describe $H(X)$ (in warm fuzzy terms) as the "uncertainty about X on average". We can also interpret it as "how random" X is. For example, variance is another quantity that describes the "randomness" in X.

The interpretation of $H(X)$ as "uncertainty" is justified by the source coding theorem, which says that $H(X)$ is the number of bits I need to describe X on average. Why is this a good measure of uncertainty?

Theorem 10.3: Source Coding Theorem

For any $\varepsilon > 0$, the iid discrete random variables $X_1, X_2, \cdots X_n \sim_{IID} P_X$ can be losslessly represented using $\leq n(H(X) + \varepsilon)$ bits (for all n sufficiently large).

Conversely, any representation using $< nH(X)$ bits is impossible without loss of information.

This result has two parts:

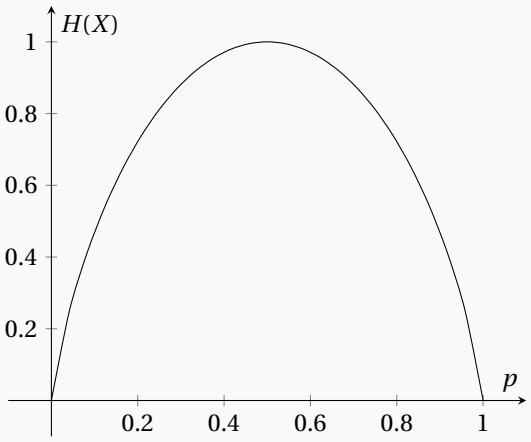
1. Descriptions $\leq n(H(X) + \varepsilon)$ bits are possible
2. Descriptions $< nH(X)$ bits are not

Example 10.4: Huffman Codes

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

Example 10.5

For $X \sim$ Bernoulli(p), we have



We can think of X as a coin flip.

- If I flip a fair coin n times, then we need n bits to represent all n outcomes.

  This makes sense, because nothing about the ith flip tells us about anything else.
- If I flip a biased ($p = 0.11$) coin n times, then I only need $\approx \frac{n}{2}$ bits to describe all n outcomes.
- If I flip a biased ($p = 0$) coin n times, then I need 0 bits to represent all n outcomes.

  This makes sense, because we always know that it'll be tails.

The intuition of the symmetry in the graph is because we don't care about what the labels of the outcomes are—we can switch heads and tails and we should get the same result.

How can the second point be possible? Concentration; from a lot of randomness comes determinism.

For a sequence $X_1, X_2, \ldots, X_n$, let the probability of observing it be:

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

In other words, all X_i's are $\sim_{IID} P_X$.

Theorem 10.6: Asymptotic Equipartition Theorem (AEP)

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

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

In other words, with overwhelming probability,

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

Proof. WLLN says that

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

□

2/25/2021

Lecture 11

Channel Coding Theorem

We will now use AEP to prove "achievability" part of the source coding theorem:

Proof. Fix $\varepsilon > 0$, and for each $n \geq 1$ define a "typical set":

$$\begin{aligned} A_\varepsilon^{(n)} &= \{(X_1, \ldots, X_n) : \mathbb{P}(X_1, \ldots, X_n) \geq 2^{-nH(X)+\varepsilon}\} \\ &= \text{a subset of possible observed sequences} \end{aligned}$$

Some properties of this set (we claim):

- $\mathbb{P}\Big((X_1, \ldots, X_n) \in A_\varepsilon^{(n)}\Big) \to 1$ as $n \to \infty$

  This is evident by AEP:

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

  since the LHS at the end is approaching $H(X)$.

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

  This comes from the definition:

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

Suppose I have N objects. How many bits (maximum) do I need to represent each object? We need $\log N$ bits (there are a total of 2^x total bitstrings, and we can just assign one to each object).

This gives us a protocol for source coding:

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

For this protocol, what is the average description length?

$$\begin{aligned} \mathbb{E}\big[\text{number of bits in description}\big] &\leq n\left(H(X) + \frac{\varepsilon}{2}\right) \underbrace{\mathbb{P}\Big((X_1, \ldots, X_n) \in A_{\varepsilon/2}^{(n)}\Big)}_{\leq 1} + n\log|X| \underbrace{\mathbb{P}\big((X_1, \ldots, X_n) \notin A_\varepsilon^{(n)}\big)}_{\leq \varepsilon/2 \text{ for } n \text{ sufficiently large}} \\ &\leq n(H(X) + \varepsilon) \text{ for all } n \text{ sufficiently large} \end{aligned}$$

This concludes the proof of achievability; the converse proof is omitted. □

11.1 Information Transmission (Channel Coding)

The main question that this addresses is: how do we send information *reliably* over an *unreliable* channel?

Let's think about an abstract communication system. Suppose we fix a "rate" $R > 0$. Communication consists of sending a message from point A to point B.

A message $M \sim \text{Uniform}\left(\{1,\dots,2^{nR}\}\right)$ takes nR bits to represent, because $H(M) = nR$.

This message goes through an encoder, which takes in the message as input, and outputs some sequence $X^n(M)$ (superscript represents a vector), i.e. $(X_1(M),\dots,X_n(M))$.

This sequence goes through a noisy channel, which corrupts the message into the sequence Y^n, i.e. $(Y_1,\dots,Y_n)$. Here, Y_i is a "noisy version" of X_i.

At the other end, we have a decoder, which takes in the noisy sequence and gets back $\hat{M}(Y^n)$.

As a diagram, this is

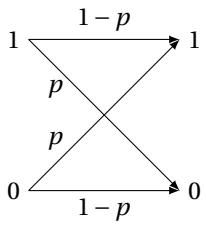
$$M \xrightarrow{\text{encoder}} X^n(M) \xrightarrow{\text{noisy channel}} Y^n \xrightarrow{\text{decoder}} \hat{M}(Y^n)$$

The parameters for this system are:

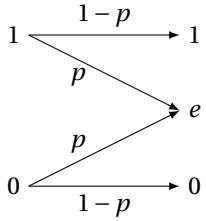
- "Rate" $= R = \frac{H(M)}{n} = \frac{\#\text{ of "information bits"}}{\#\text{ of channel uses}}$
- "error probability" $= P_e^{(n)} = \mathbb{P}\left(\hat{M} \neq M\right)$

Examples of "noisy channels":

- Binary symmetric channel $BSC(p)$: flips a bit independently with probability p



- Binary erasure channel $BEC(p)$: erases a bit independently with probability p



Channels in general are represented by a conditional pmf $P_{Y|X}$.

For example, the $BSC(p)$ channel is represented by

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

Definition 11.1: Mutual Information

For a channel $P_{Y|X}$ and input distribution P_X, and there is a joint distribution $P_{X,Y}(x,y) = P_X(x)P_{Y|X}(y \mid x)$ and a marginal distribution of outputs $P_Y(y) = \sum_x P_{X,Y}(x,y)$, we have the mutual information

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

Definition 11.2: Channel Capacity

For a channel $P_{Y|X}$, the capacity is defined as

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

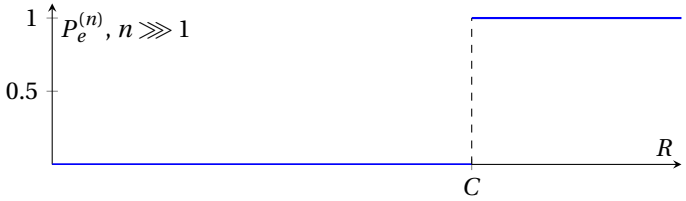
In other words, it's the maximum mutual information between channel input and output over all input distributions. Note that there is no dependence on n.

Theorem 11.3: Shannon's Channel Coding Theorem

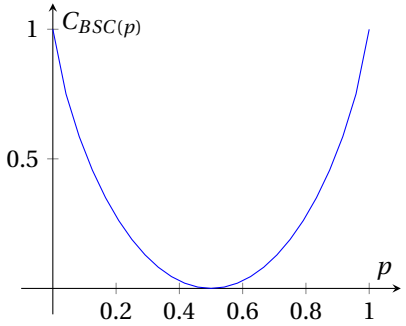
Fix a channel $P_{Y|X}$ and $\varepsilon > 0$ and $R < C$.

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

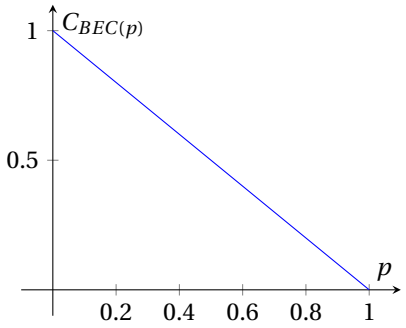
What does this mean? We essentially have this graph of error probability:



For example, for a $BSC(p)$ channel, $C = 1 - H_2(p) = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$, i.e.



For a $BEC(p)$ channel, $C = 1 - p$, i.e.



Our next step is to prove the channel coding theorem for the special case of $BEC(p)$.

We need to show two things:

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

To prove the first point:

Proof. Consider a block of n channel uses.

The transmitter doesn't know what locations will be erased. But, let's suppose a genie told us. This means that a transmitter can send information without error in unerased positions.

$$0\ 1\ \mathrm{e}\ \mathrm{e}\ 1\ \mathrm{e} \cdots 0\ \mathrm{e}$$

How many unerased positions are there? $< n(1-p+\varepsilon)$ with overwhelming probability for any $\varepsilon > 0$ and all n sufficiently large.

This tells us that the transmitter can only reliably send $\approx n\big(1-p\big)$ bits, and therefore $R \leq (1-p)$. □

To prove the second point:

Proof. This relies on something called the "probabilistic method". This refers to the general strategy to show the existence of something you want. Specifically, don't actually construct an explicit scheme and analyze it; just show one exists with probability > 0.

Fix $\varepsilon > 0$, and fix $R < 1 - p - \varepsilon$. Generate a random $2^{nR} \times n$ matrix (a "codebook")

$$\mathbf{C} = \begin{bmatrix} C_{11} & \cdots & C_{1,n} \\ \vdots & \ddots & \vdots \\ C_{2^{nR},1} & \cdots & C_{2^{nR},n} \end{bmatrix}$$

where $C_{ij} \sim_{IID} \text{Bernoulli}\left(\frac{1}{2}\right)$.

Give $\mathbf{C}$ to both the encoder and the decoder. The protocol is thus

- On observing the message $M \in \{1, \ldots, 2^{nR}\}$, send the row M of C.
- On receiving Y^n, look for row in C that matches (modulo erasures). In other words, find a row that matches all the unerased packets; the index of the row is the message packet.

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

3/2/2021

Lecture 12

Markov Chains

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

Let $E = \{1, 2, \ldots, n\}$ (the set will be abbreviated as $[n]$) be the positions that are erased by the channel.

WLOG (by symmetry) I can assume $M = 1$ is correct.

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

The upper bound basically splits the sum into two groups based on the size of E, but we upper bound the cases of $|E| > n(p + \frac{\varepsilon}{2})$. That last term must be vanishing as $n \to \infty$ by WLLN (if we divide by n on both sides of the inequality).

So, let's just look at $\mathbb{E}\left[\mathbf{1}\{\hat{M} \neq M\} \mid E\right]$ for $|E| \leq n\left(p + \frac{\varepsilon}{2}\right)$:

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

What this RHS means is that we know that our received message will match with row 1, disregarding the erasures (E). Errors occur if any other row (m) matches row 1.

Now, we can use the union bound, noting that the probability that a row matches is just the probability that the $n - |E|$ non-erased bits match (with probability $\frac{1}{2}$):

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

In the last upper bound, we note that $nR - n + |E| \leq n\left(1 - p - \frac{\varepsilon}{2}\right) - n + n\left(p + \frac{\varepsilon}{2}\right) = -n\varepsilon$.

Going back to what we had before, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{C}}[P_e^{(n)}] &\leq \sum_{E : |E| \leq n(p+\frac{\varepsilon}{2})} 2^{-n\varepsilon} P(E) + \mathbb{P}\left(\frac{1}{n}|E| > p + \frac{\varepsilon}{2}\right) \\ &\leq 2^{-n\varepsilon} + \mathbb{P}\left(\frac{1}{n}|E| > p + \frac{\varepsilon}{2}\right) \\ &\to 0 \text{ as } n \to \infty\end{aligned}$$

This means that there must exist some choice of codebook and n sufficiently large to make $P_e^{(n)} < \varepsilon$.

12.1 Markov Chains

Random variables by themselves are only so interesting; often, we are interested in sequences of random variables, for example $(X_n)_{n \geq 0}$, to model real-life things. These are called random processes or stochastic processes. (You just use "stochastic" if you're feeling fancy.)

For example, these could model

- Robot position over time
- Website visited by internet user
- Signal received by cell tower
- etc.

All of these things have some temporal aspect.

Up to now, the only processes we've seen have been IID. This is a good starting point with nice results (WLLN, CLT, etc.), but are limited for modeling real scenarios.

Markov Chains are one level above IID processes. They're simple enough that we can say some very strong quantitative things about them. They're a very flexible class of processes, and useful for modeling a wide variety of situations.

Definition 12.1: Markov Chain

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

$$\mathbb{P}\big(X_{n+1} = j \mid X_n = i, X_{n-1} = x_{n-1}, \cdots, X_0 = x_0\big) = \mathbb{P}\big(X_{n+1} = j \mid X_n = i\big).$$

In other words, the future only depends on the past through the present—the next state depends only on the current state, and not on anything before.

An interesting note is that any process with finite memory is equivalent to this Markov process; the state space would just be augmented as $\mathcal{S}^w$ if we have w steps of memory.

Some terminology:

Definition 12.2: State

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

If $(X_n)_{n\geq 0}$ is a MC, then there is an implicit underlying probability space $(\Omega, \mathcal{F}, P)$ on which X_n's are all random variables.

Definition 12.3: Temporally Homogeneous Markov Chain

Temporally Homogeneous Markov Chains are MC such that

$$\mathbb{P}\big(X_{n+1} = j \mid X_n = i\big) = p_{ij},$$

for $\forall i, j \in \mathcal{S}$, $n \geq 0$. These p_{ij} are "transition probabilities" from i to J, not dependent on time.

In this class, we'll only be working with temporally homogeneous Markov chains (and we'll be calling them Markov chains without the qualifier).

Note that these transition probabilities have to satisfy a few rules:

- $p_{ij} \geq 0$, $\forall i, j \in \mathcal{S}$
- $\sum_{j\in\mathcal{S}} p_{ij} = 1$, $\forall i \in \mathcal{S}$

  Why? This is just the sum of the probabilities for all possible next states, which is just 1.

The transition probabilities p_{ij} for $i, j \in \mathcal{S}$ describe the statistics of the Markov chain.

It's often helpful to collect these into a transition matrix such that $[\mathbf{P}]_{ij} = p_{ij}$:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots \\ p_{21} & p_{12} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}.$$

This is a "stochastic matrix"; the entries are nonnegative, and the rows sum to 1.

MCs can be represented by state transition diagrams:

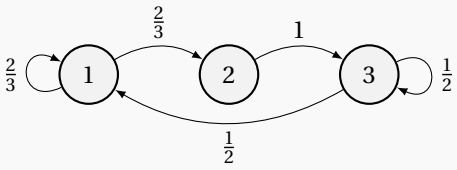
- Each state is represented by a node
- Arrows between states with transition probabilities. (No arrow if there is no transition probability)

Example 12.4

If we have

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

The state transition diagram would be



The transition matrix $\mathbf{P}$ tells us "one-step" transition probabilities.

What about $\mathbb{P}\big(X = j \mid X_0 = i\big)$ (i.e. the "n-step transition probability")? This is by definition p_{ij}^n. (*not* some quantity p_{ij} raised to the nth power!)

Theorem 12.5: Chapman-Kolmogorov Equations

n-step transition probabilities can be computed as $p_{ij}^n = [\mathbf{P}^n]_{ij}$ (the latter *is* raising $\mathbf{P}$ to the nth power).

Proof. Induct on n. The base case $n = 1$ holds by definition.

$$\begin{aligned}
\mathbb{P}\big(X_{n+1} = j \mid X_0 = j\big) &= \sum_{k \in \mathcal{S}} \mathbb{P}\big(X_{n+1} = j, X_n = k \mid X_0 = i\big) \\
&= \sum_{k \in \mathcal{S}} \mathbb{P}\big(X_{n+1} = j \mid X_n = k, \cancel{X_0 = i}\big) \mathbb{P}(X_n = k \mid X_0 = i) \\
&= \sum_{k \in \mathcal{S}} \big\langle j\text{th column of } \mathbf{P}, i\text{th row of } \mathbf{P}^n \big\rangle \\
&= [\mathbf{P}^n \times P]_{ij} = [\mathbf{P}^{n+1}]_{ij}
\end{aligned}$$

□

3/4/2021

Lecture 13

Big Theorem for DTMCs

13.1 Classification of States

If there is a path in the state-transition diagram from i to j (i.e. $p_{ij}^n > 0$ for some $n \geq 1$), then we say that j is accessible from i, and we write $i \to j$.

If we also have that $j \to i$, then we write $i \leftrightarrow j$, and states i, j "communicate".

By convention, we say $i \leftrightarrow i$, $\forall i \in \mathcal{S}$.

Our claim is that $\leftrightarrow$ is an equivalence relation on $\mathcal{S}$. That is,

1. $i \leftrightarrow i$, $\forall i \in \mathcal{S}$
2. $i \leftrightarrow j \iff j \leftrightarrow i$, $\forall i, j \in \mathcal{S}$

3. $i \leftrightarrow k, k \leftrightarrow j \implies i \leftrightarrow j, \forall i, j, k \in \mathcal{S}$

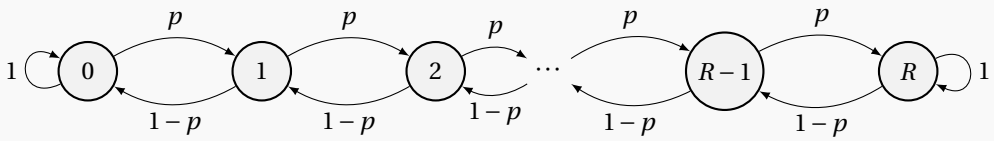
What does an equivalence relation do? Equivalence relations on a set partition the set into "equivalence classes".

So, $\leftrightarrow$ partitions $\mathcal{S}$ into "classes"' of communicating states.

If $C \subset \mathcal{S}$ is a "class" and $i \in C$, then $j \in C \iff i \leftrightarrow j$.

Example 13.1: Gambler's Ruin

Suppose we fix $p \in (0,1)$.



What classes of states is there?

$$\{0\}, \{1, 2, \cdots, R-1\}, \{R\}.$$

Definition 13.2: Irreducibility

A MC is *irreducible* if it has only one class (i.e. $\mathcal{S}$).

13.2 Class Properties

Definition 13.3: Recurrence

A state $i \in \mathcal{S}$ is said to be *recurrent* if, given that $X_0 = i$, the process revisits state i with probability 1.

This is equivalent to saying that I will visit state i infinitely many times with probability 1, given that I start in state i. For example, in Gambler's Ruin, states 0 and R are recurrent.

Definition 13.4: Transience

A state $i \in \mathcal{S}$ is *transient* if it is not recurrent.

This is equivalent to saying that if $X_0 = i$, then $(X_n)_{n \geq 1}$ will visit i finitely many times with probability 1. For example, in Gambler's Ruin, states $\{1, \ldots, R-1\}$ are transient.

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

Proof. It suffices to show if i is recurrent, then so is j. In particular, it suffices to show that if $X_0 = i$, then I'll land in j after finite time with probability 1.

Since $i \leftrightarrow j$, $\exists n \geq 1$ s.t. $p_{ij}^n > 0$. So, I'll land in j after $\text{Geom}(p_{ij}^n)$ visits to I. □

We define

$$T_i = \min\{n \geq 1 : X_n = i\}.$$

In other words, T_i is the first time ($n \geq 1$) that I enter state I.

Definition 13.5: Positive and Null Recurrence

If $i \in \mathcal{S}$ is recurrent, we further classify

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

Positive and null recurrence are also class properties.

Definition 13.6: Periodicity

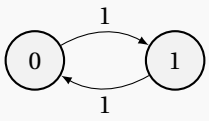
For $i \in \mathcal{S}$, let

$$\text{period}(i) = \gcd\{n \geq 1 : p_{ii}^n > 0\}.$$

In words, if I start in state i, then revisits to state i only occur at integer multiples of the period period(i).

Example 13.7

This is a periodic MC:



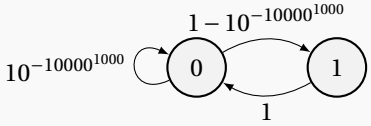
Periodicity is also a class property. In other words, if $i \leftrightarrow j$, then period(i) = period(j).

Definition 13.8: Aperiodicity

An irreducible MC is *aperiodic* if any state (and therefore all states) has period 1.

Example 13.9

A remark: periodicity is a "brittle" property; it's not very robust. If we just added one self loop to the previous example, then it becomes aperiodic, although it won't realistically get to that self loop even if we simulated this until the heat death of the universe.



Definition 13.10: Stationary Distribution

A probability distribution $\pi = (\pi_i)_{i \in \mathcal{S}}$, i.e. a row vector, is said to be a *stationary distribution* if $\pi = \pi \mathbf{P}$. Written out, this is just saying that $\pi_j = \sum_{i \in \mathcal{S}} \pi_i p_{ij}$, $\forall j \in \mathcal{S}$.

This is also a left eigenvector of $\mathbf{P}$ with nonnegative entries that sum to 1.

The idea here is that if $X_0 \sim \pi$, then $X_n \sim \pi$, $\forall n \geq 0$. In other words, the distribution over states is invariant in time, and the resulting process $(X_n)_{n \geq 0}$ is "stationary", i.e. not changing with time.

Theorem 13.11: "Big Theorem" For Markov Chains

(This is essentially a statement about long-term behavior.)

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

1. Either all states are transient, or all are null recurrent.

   In this case, no stationary distribution exists, and

   $$\lim_{n\to\infty} p_{ij}^n = 0 \quad \forall i,j \in \mathcal{S}.$$

2. All states are positive recurrent.

   In this case, a stationary distribution π exists. It is unique and satisfies

   $$\pi_j = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} p_{ij}^k = \frac{1}{\mathbb{E}\left[T_j \mid X_0 = j\right]} \quad \forall i,j \in \mathcal{S}.$$

   Moreover, if the MC is aperiodic, then

   $$\lim_{n\to\infty} p_{ij}^n = \pi_j \quad \forall i,j \in \mathcal{S}.$$

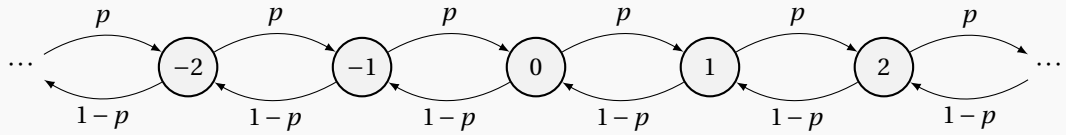
We pretty much always know the long-term behavior of irreducible MCs. As an explanation,

1. If all states are transient, then we return to the state finitely many times with probability 1, so the probability that we're at a given state at time n goes to 0, regardless of where we start.
2. In this second case, the aperiodic case converges in distribution to the stationary distribution.

Remark: every irreducible finite-state MC is positive recurrent. By the pigeonhole principle, we have to return to a state with positive probability for at least one state. That is, if we have a finite-state MC, then only (2) is possible.

Example 13.12

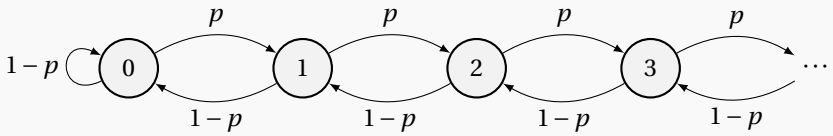
Consider a random walk on $\mathbb{Z}$ with transition probability $p \in (0,1)$. That is,



This chain is irreducible (all states communicate with each other). If $p = \frac{1}{2}$, then all states are null recurrent. That is, by CLT we will always return to the starting state with probability 1 (it'll be some Gaussian).

If $p \neq \frac{1}{2}$, then all states are transient (by SLLN). This is because we'll always be drifting either to the right or to the left.

Example 13.13: "Birth-Death" Chain



This chain is positive recurrent if $p < \frac{1}{2}$, null recurrent if $p = \frac{1}{2}$, and transient if $p > \frac{1}{2}$.

Example 13.14: Page Rank

We model the internet by a connected finite directed graph $G = (V, E)$. Let $(X_n)_{n\geq 0}$ be a random walk on this graph, where the next state is chosen uniformly from outgoing links.

The main question is: how do we rank importance of webpages?

By the Big Theorem, a stationary distribution for this process exists for this MC, and π_j describes the fraction of traffic on site j in the long run. So, π_j is a good proxy for the "rank" of page j.

3/9/2021

Lecture 14

Reversibility, First Step Equations

Example 14.1

Let $(X_n)_{n\geq 0}$ be a random walk on the hypercube $\{0,1\}^n$, where the next vertex is chosen by randomly flipping one of the bits of the current state.

If we start at state $(0,\cdots,0)$, what is the expected number of steps before returning?

We know that all states are positive recurrent, so by the Big Theorem, $\mathbb{E}[T_{0\cdots 0} \mid X_0 = (0\cdots 0)] = \frac{1}{\pi_{0\cdots 0}} = \frac{1}{2^{-n}} = 2^n$. This is because the graph is symmetric, so the stationary distribution must be uniform.

Example 14.2

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

$$\underbrace{\frac{1}{n}\sum_{k=1}^{n} R(X_k)}_{\text{avg reward in first } n \text{ steps}} \xrightarrow{\text{a.s.}} \mathbb{E}[R(x)],\ X \sim \pi.$$

That is, the empirical average of rewards tends to the true mean of rewards at the stationary distribution. Why is this the case?

Let $N_j(n)$ be the number of entries into j up to time n. This means we can write

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

Each re-entry time is iid, so the $\frac{N_j(n)}{n}$ is the empirical average of re-entry times. This allows us to use SLLN to simplify it into an expectation.

14.1 Reversibility

How do we compute the stationary distribution?

The general answer unfortunately is to solve $\pi_j = \sum_j \pi_i p_{ij}, \forall j$. It turns out this is much easier when the MC is "reversible".

Definition 14.3: Reversibility

An irreducible irreducible MC is *reversible* if there exists a probability vector π satisfying

$$\pi_j P_{ji} = \pi_i P_{ij} \quad \forall i, j \in \mathcal{S}.$$

These equations are called the *detailed balance equations*.

We can think of the DBEs as a notion of "flow"; the amount of mass flowing from j to i is the same as the amount of mass flowing from i to j.

Lemma 14.4

If a MC is reversible, then π is a stationary distribution. (In fact, unique by irreducibility + Big Theorem.)

Proof.

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

This equation is just the definition of the stationary distribution. □

So, in the case of a reversible MC, just solve detailed balance equations for π. Sometimes we can *assume* the MC is reversible, and try solve the DBEs; if it works out, we've found the stationary distribution and deduced that it's reversible. Otherwise, we'd need to solve the original equations with eigenvalues/eigenvectors.

Where does the term "reversible" come from? If we start a reversible MC with $X_0 \sim \pi$, then the sequence of states $(X_0, X_1, \cdots, X_n) \overset{d}{=} (X_n, X_{n-1}, \cdots, X_0)$, i.e. the two sequences are equal in distribution.

In other words, reversing time on a reversible MC would look exactly the same as if we went normally. This is like equilibrium in physical systems.

The Big Theorem tells us about the asymptotic behavior of an irreducible MC. Now, let's turn our attention to techniques for analyzing finite-horizon (short-term) behavior of (not necessarily irreducible) MCs.

14.2 First Step Analysis

14.2.1 Hitting Times

Definition 14.5: Hitting time

Consider a subset of states $A \subset \mathcal{S}$; we define the *hitting time* to be

$$T_A = \min\{n \geq 0 : X_n \in A\}.$$

T_A is a random variable, but trying to compute the distribution of the hitting time can be incredibly complicated; but it does make sense to look at the expectation, and is much easier.

Here's our strategy. Let us define some quantities $t_i = \mathbb{E}[T_A \mid X_0 = i]$. Now, we formulate the "first step" equations, i.e. a way of thinking recursively about the hitting time in terms of how the MC process proceeds.

That is, for $i \neq A$, the soonest we can hit A is on the next step. In other words,

$$\mathbb{E}[T_A \mid X_0 = i] = 1 + \sum_{j \in \mathcal{S}} p_{ij} \mathbb{E}\big[T_A \mid X_0 = j\big].$$

This should make sense, because at the next transition, we take one step, and once we're at state j, it's just like we started in state j again.

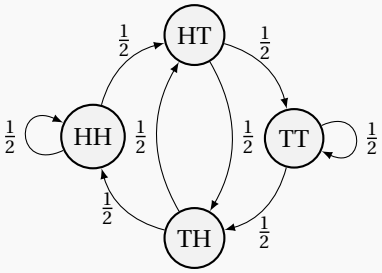
Additionally, for $i \in A$, we have that $\mathbb{E}[T_A \mid X_0 = i] = 0$.

This is just a system of equations, called the first step equations (FSE);

$$\begin{cases} t_i = 1 + \sum_j p_{ij} t_j & i \notin A \\ t_i = 0 & i \in A \end{cases}.$$

Example 14.6

Consider fair coin tosses. How many tosses do we need until we get two tails in a row?



If we label the nodes in order HH, HT, TH, TT, then our transition matrix is

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

To answer the question, we can set up FSE, starting in state HH, i.e. we compute $t_{HH} = \mathbb{E}[T_{TT} \mid X_0 = HH]$.

$$\begin{aligned} t_{HH} &= 1 + \frac{1}{2} t_{HH} + \frac{1}{2} t_{HT} \\ t_{HT} &= 1 + \frac{1}{2} t_{HT} + \frac{1}{2} t_{TT} \\ t_{TH} &= 1 + \frac{1}{2} t_{TH} + \frac{1}{2} t_{TT} \\ t_{TT} &= 0 \end{aligned}$$

Solving this system gives us $t_{HH} = 6$, $t_{HT} = 4$, and $t_{TH} = 6$. This means that it takes us 6 tosses on average to get 2 tails in a row.

3/11/2021

Lecture 15

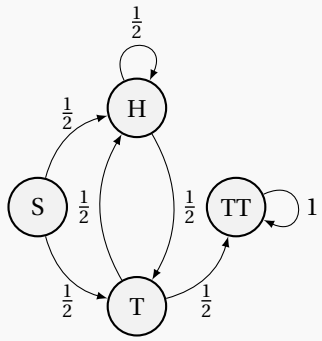
Poisson Processes

Let's continue on with the first step equation examples from last lecture.

15.0.1 Collected Rewards

Example 15.1

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



Suppose $t_i = \mathbb{E}[\#\text{ heads seen} \mid X_0 = i]$, for $i \in \mathcal{S}$. Setting up a system of equations, we have

$$\begin{aligned} t_S &= \frac{1}{2}t_H + \frac{1}{2}t_T \\ t_H &= 1 + \frac{1}{2}t_H + \frac{1}{2}t_T \\ t_T &= \frac{1}{2}t_H + \frac{1}{2}t_{TT} \\ t_{TT} &= 0 \end{aligned}$$

Solving this system, we have that $t_S = 3$, $t_H = 4$, and $t_T = 2$. Therefore, we expect to see 3 heads, starting at the start state.

15.0.2 Hitting Probabilities

Consider states $A, B \subset \mathcal{S}$, $A \cap B = \varnothing$, and $T_A = \min\{n \geq 0 : X_n \in A\}$, $T_B = \min\{n \geq 0 : X_n \in B\}$.

Definition 15.2: Hitting Probability

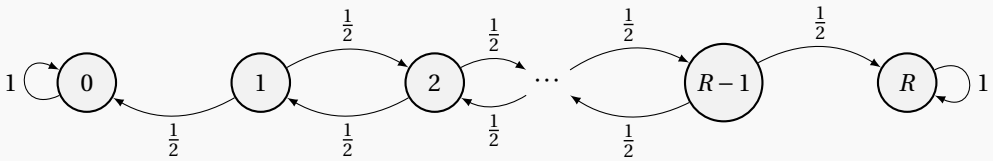
The "hitting probability" is the quantity $\mathbb{P}(T_A < T_B \mid X_0 = i)$.

The way to solve these kinds of problems is to define a system of equations $\alpha(i) := \mathbb{P}(T_A < T_B \mid X_0 = i)$, for $i \notin A \cup B$. Trivially, $\alpha(i) = 0$ for $i \in B$, and $\alpha(i) = 1$ for $i \in A$.

For $i \notin A \cup B$, we have equivalently (by law of total probability and the Markov Property)

$$\mathbb{P}(T_A < T_B \mid X_0 = i) = \sum_{j \in \mathcal{S}} p_{ij}\, \mathbb{P}\big(T_A < T_B \mid X_0 = j\big).$$

Example 15.3: Gambler's Ruin



Our claim is that $\mathbb{P}(T_R < T_0 \mid X_0 = i) = \frac{i}{R}$ for $i = 0 \cdots R$.

To verify, we find that

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

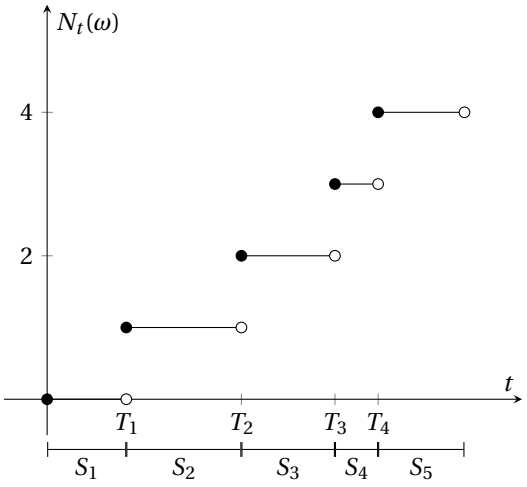
15.1 Poisson Processes

A Poisson Process is an (*the*) example of a "Counting Process". It forms the basis for continuous time MCs.

Definition 15.4: Counting Process

A counting process is a sequence of random variables $(N_t)_{t \geq 0}$ (here, $t \in \mathbb{R}$) is a continuous-time integer-valued random process, which has right-continuous sample paths.

In a picture, we have



The "arrival times" $T_i = \min\{t \geq 0 : N_t \geq i\}$, or the time of the ith arrival.

The interarrival times $S_i = T_i - T_{i-1}$ for $i \geq 1$ are the durations between arrivals.

Definition 15.5: Poisson Process

A rate-λ Poisson Process is a counting process with iid interarrival times $S_i \sim_{IID} \text{Exp}(\lambda)$.

Equivalently, the following three conditions must all hold:

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

Why "Poisson"?

Theorem 15.6: Poisson Distribution of Point Counts

If $(N_t)_{t\geq 0}$ is a PP(λ), then for each $t \geq 0$, $N_t \sim \text{Pois}(\lambda t)$.

In other words, $\mathbb{P}(N_t = n) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$, which is the Poisson pmf.

Proof.

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

In the last equality, we use the fact that $S_{n+1} \sim \text{Exp}(\lambda)$.

Fact: the sum of n independent exponentials is an "Erlang" RV. We can use this density here.

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

□

From here, we can make a few observations.

By the memoryless property of Exp(λ), if we have a $(N_t)_{t\geq 0} \sim \text{PP}(\lambda)$, then $(N_{t+s} - N_s)_{t\geq 0}$ is also a PP(λ), for all $s \geq 0$. Moreover, $(N_{t+s} - N_s)_{t\geq 0}$ is independent of $(N_\tau)_{0\leq \tau \leq s}$. In particular, Poisson processes have a Markov Property.

Poisson processes have independent and stationary increments. In other words, if $t_0 < t_1 < t_2 < \cdots < t_k$, then the increments $(N_{t_1} - N_{t_0}), (N_{t_2} - N_{t_1}), \cdots, (N_{t_k} - N_{t_{k-1}})$ are independent, and $(N_{t_i} - N_{t_{i-1}}) \sim \text{Pois}(\lambda(t_i - t_{i-1}))$, $\forall i$.

15.2 Conditional Distribution of Arrivals

Theorem 15.7: Uniform Arrival Times

Conditioned on the event $\{N_t = n\}$, we have

$$(T_1, T_2, \cdots, T_n) \overset{d}{=} (U_{(1)}, U_{(2)}, \cdots, U_{(n)}),$$

where $U_{(i)}$ are order statistics of n Uniform$(0, t)$ RVs.

In other words, given n arrivals occurred up to time t, the arrival times look like iid Uniform$(0, t)$ RVs in distribution.

Proof. For $0 = t_0 \leq t_1 \leq \cdots \leq t_n \leq t$, we have

$$f_{T_1 T_2 \cdots T_n | N_t}(t_1, \cdots t_n \mid n) = \frac{\mathbb{P}(N_t = n \mid T_1 = t_1, \cdots, T_n = t_n)}{\mathbb{P}(N_t = n)} f_{T_1 \cdots T_n}(t_1, \cdots, t_n)$$

$$= \frac{\mathbb{P}\left(N_t - N_{t_n} = 0 \mid \cancel{T_1 = t_1, \cdots T_n = t_n}\right)}{\mathbb{P}(N_t = n)} \prod_{i=1}^{n} f_{S_i}(t_i - t_{i-1})$$
$$= \frac{e^{-\lambda(t-t_n)}}{e^{-\lambda t}\frac{(\lambda t)^n}{n!}} \prod_{i=1}^{n} \lambda e^{-\lambda(t_i - t_{i-1})}$$
$$= \frac{n!}{t^n} \text{ after cancellation}$$

This last expression is precisely the density of order statistics of $(U_{(1)}, \cdots, U_{(n)})$.

Why? Recall that uniform distributions on $(0,t)^n$ has density $\frac{1}{t^n}$ on $[0,t]^n$. Order statistics sort among $n!$ permutations, so we get a $n!$ multiplier on the region where coordinates are sorted. □

Example 15.8

Suppose cars pass through a toll booth according to $PP(\lambda)$, where λ = vehicles/minute.

We can do a lot of things since this is a Poisson process:

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

$$\mathbb{P}(N_2 = 0) = e^{-2\lambda}.$$

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

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

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

   $\frac{10}{4}$ because the 10 vehicles' arrival times are uniformly distributed over the 2 minutes.

3/17/2021

Lecture 16

Poisson Merging/Splitting, Continuous Time Markov Chains

Example 16.1

Suppose photons arrive at a detector $\sim$ PP(λ). If 10^6 photons are detected in 2 seconds, what is the distribution of the number of photons that arrived in the first second?

Since the (ordered) arrival times are uniform over the 2 second interval, each of the 10^6 photons has a $\frac{1}{2}$ probability of being in the first second. This means that the number of photons in the first second $\sim \text{Bin}(10^6, \frac{1}{2})$.

16.1 Poisson Merging and Splitting

Theorem 16.2: Poisson Merging

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

Proof. Let's go through the three conditions for Poisson processes:

1. We have that $N_{1,0} + N_{2,0} = 0 + 0 = 0$
2. We also have that

$$\begin{aligned}(N_{1,t} + N_{2,t}) - (N_{1,s} + N_{2,s}) &= (N_{1,t} - N_{1,s}) + (N_{2,t} - N_{2,s}) \\ &\overset{d}{=} \text{Pois}(\lambda_1(t-s)) * \text{Pois}(\lambda_2(t-s)) \\ &= \text{Pois}((\lambda_1 + \lambda_2)(t-s))\end{aligned}$$

3. Finally, $(N_{1,t} + N_{2,t})_{t\geq 0}$ has independent increments. This is true because both $(N_{1,t})_{t\geq 0}$ and $(N_{2,t})_{t\geq 0}$ have independent increments.

Since all three properties are true, $(N_{1,t} + N_{2,t})$ must also be a Poisson process $\text{PP}(\lambda_1 + \lambda_2)$. □

Example 16.3

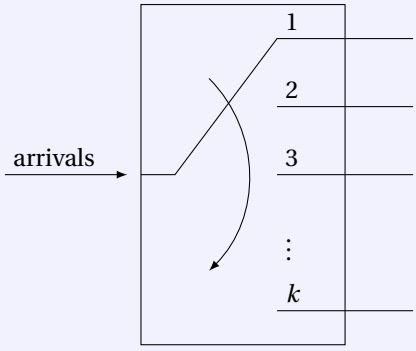
Suppose men are hospitalized $\sim \text{PP}(\lambda_1)$, independent of women who are hospitalized $\sim \text{PP}(\lambda_2)$.

Then, the total number of men and women hospitalized $\sim \text{PP}(\lambda_1 + \lambda_2)$.

Theorem 16.4: Poisson Splitting/Thinning

Let $p_1, \cdots, p_k$ be nonnegative, with $\sum_{i=1}^k p_i = 1$. Let $(N_t)_{t\geq 0}$ be a $\text{PP}(\lambda)$. We "mark" each arrival with a label "i" with probability p_i, independently of all other arrivals. Let $(N_{i,t})_{t\geq 0}$ be the process that counts arrivals marked with "i", for $i = 1, \ldots, k$.

We can visualize this as a random switch in position i with probability p_i. This splits up the arrivals into k groups, one for each mark.



Note that $N_t = \sum_{i=1}^k N_{i,t}$ for all $t \geq 0$.

We have that $(N_{i,t})_{t\geq 0}$ for $i = 1, \ldots, k$ are independent poisson processes with respective rates $p_i\lambda$.

Proof. We can consider $k = 2$ (because we can keep subdividing). Here, we have $p_1 = p$ and $p_2 = 1 - p$.

$$\begin{aligned}\mathbb{P}\big(N_{1,t}=n, N_{2,t}=m\big) &= \mathbb{P}\big(N_{1,t}=n, N_{2,t}=m, N_t=n+m\big)\\ &= \mathbb{P}\big(N_{1,t}=n, N_{2,t}=m \mid N_t=n+m\big)\mathbb{P}(N_t=n+m)\\ &= \binom{n+m}{n} p^n (1-p)^m e^{-\lambda t}\frac{(\lambda t)^{n+m}}{(n+m)!}\\ &= e^{-p\lambda t}\frac{(p\lambda t)^n}{n!}\cdot e^{-(1-p)\lambda t}\frac{\big((1-p)\lambda t\big)^m}{m!}\end{aligned}$$

These are the pmfs are Pois$(p\lambda t)$ and Pois$((1-p)\lambda t)$.

Showing independence is a little more involved but the proofs are not shown here. □

Example 16.5

Suppose packets arrive to a router ~ PP(λ). They are randomly routed to outgoing link A with probability p, and routed to outgoing link B with probability $1-p$.

Packets on link A ~ PP$(p\lambda)$, and packets on link B ~ PP$((1-p)\lambda)$, and are independent.

These properties seem simple, but they can be very powerful for problem solving.

Example 16.6: Random Incidence Paradox

Consider $(N_t)_{t\geq 0}$ ~ PP(λ). Suppose I pick a "random" time t_0. What is the expected length of the interarrival interval in which t_0 falls?

Conventional wisdom would say $\frac{1}{\lambda}$ = expected interarrival time.

Say t_0 falls between arrivals T_i and T_{i+1}. The length of the interarrival time we arrive in is

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

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

Let's look at the complementary cdf:

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

This means that $t_0 - T_i \sim \text{Exp}(\lambda)$. By linearity of expectation,

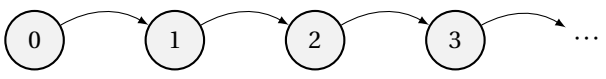
$$\mathbb{E}[L] = \frac{2}{\lambda}.$$

This turns out to be twice the average interarrival time!

What's the explanation for this? If we show up at a random time, we're more likely to show up in a long interval.

16.2 Continuous Time Markov Chains

Consider representing a PP(λ) as follows:



We start in state 0, and wait for Exp(λ) amount of time before transitioning. Thus, if $(X_t)_{t\geq 0}$ is a process where X_t is the state at time $t \geq 0$, then $(X_t)_{t\geq 0}$ is PP(λ).

This is an example of a CTMC. In fact, all CTMCs look sort of like this.

Similar to DTMCs, we assume that there is a countable state space $\mathcal{S}$ for CTMCs.

Recall that DTMCs are defined by a transition matrix $\mathbf{P}$.

Definition 16.7: Rate Matrix

CTMCs are defined in terms of a *rate matrix* $\mathbf{Q}$ satisfying:

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

That is, off diagonal elements of $\mathbf{Q}$ are nonnegative, and the rows of $\mathbf{Q}$ sum to 0.

Definition 16.8: Transition Rate

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

Definition 16.9: Jump Chain

Note that $[\mathbf{Q}]_{ij} = q_i p_{ij}$ for all i, j for some $(p_{ij})_{i,j\in\mathcal{S}}$ satisfying $\sum_{j\in\mathcal{S}} p_{ij} = 1$, $p_{ii} = 0$, and $p_{ij} \geq 0$.

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

The way a CTMC with rate matrix $\mathbf{Q}$ works is as follows. We start at state $X_0 = i$, and then:

- Hold for Exp(q_i) amount of time, then jump to state j with probability p_{ij}, for $j \in \mathcal{S}$.
- Hold for Exp(q_j) amount of time, then jump to state $k \in \mathcal{S}$ with probability p_{jk}, for $k \in \mathcal{S}$.
- Repeat.

X_t is the state at time $t \geq 0$. $(X_t)_{t\geq 0}$ is a CTMC.

Why is this called a Markov Chain? By the memoryless property of exponential distributions (the hold time), we have

$$\mathbb{P}\big(X_{t+\tau} = j \mid X_t = i, X_s = i_s, 0 \leq s < t\big) = \mathbb{P}\big(X_{t+\tau} = j \mid X_t = i\big).$$

It turns out that any (ruling out pathological situations) continuous time process with the above Markov property can be realized using the above procedure.

3/18/2021

Lecture 17

CTMC Examples, Big Theorem for CTMCs

All (ish) CTMCs can be constructed as follows. The parameters are

- q_i: the "transition rate" for state $i \in \mathcal{S}$
- p_{ij}: transition probability for $i \to j$ in "jump chain"

Here, $p_{ij} \geq 0$, $p_{ii} = 0$, $\sum_{j\in\mathcal{S}} p_{ij} = 1$.

The "jump chain" is a DTMC with transition probabilities p_{ij} for $i, j \in \mathcal{S}$.

The CTMC visits the same sequence of states as the jump chain, except we hold in state i for $\text{Exp}(q_i)$ time before making a transition to the next state (where we jump to j with probability p_{ij}).

We define $p_{ii} = 0$ because we want the memoryless property of exponential distributions; if $p_{ii} \neq 0$, then we'd be staying in the state for a sum of exponentials, which is not memoryless and destroys the Markov property.

An equivalent point of view for CTMCs is as follows. We can define "jump rates" as $q_{ij} \coloneqq q_i p_{ij}$, for $i, j \in \mathcal{S}$. Upon entering state i, consider the independent RVs $T_j \sim \text{Exp}(q_{ij})$ for all $j \in \mathcal{S} \setminus \{i\}$, and jump to state $j^* = \arg\min_{j \in \mathcal{S}}(T_j : j \in \mathcal{S})$ at time T_{j^*}.

In other words, we start a bunch of timers upon entering a state, and go to the state whose timer runs out first.

All of the parameters for a CTMC can be summarized in terms of a "rate matrix" $\mathbf{Q}$, defined for $i, j \in \mathcal{S}$ as

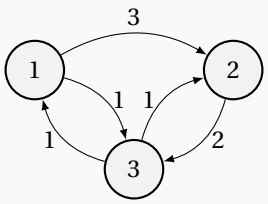
$$[\mathbf{Q}]_{ij} = \begin{cases} -q_i & i = j \\ q_{ij} & j \neq i \end{cases}.$$

Example 17.1

If $\mathbf{Q} = \begin{bmatrix} -4 & 3 & 1 \\ 0 & -2 & 2 \\ 1 & 1 & -2 \end{bmatrix}$, then we can just read off the transition rates: $q_1 = 4$, $q_2 = 2$, and $q_3 = 2$.

The jump chain then has transition probability matrix $\mathbf{P} = \begin{bmatrix} 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$.

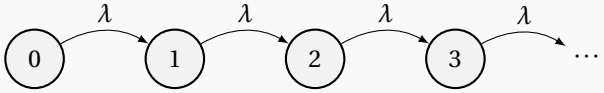
For CTMC, we draw a state-transition diagram and label transitions with jump rates:



When it comes to CTMCs, a key skill is formulating a CTMC given a description of a system model; i.e. specify jump rates q_{ij} for $i \neq j \in \mathcal{S}$.

Example 17.2

Consider a Poisson process with rate λ. The state space $\mathcal{S} = \{0, 1, 2, \ldots\} = \mathbb{N}$.



This means that the transition rates $q_{n,n+1} = \lambda$ for $n \geq 0$, and $q_{ij} = 0$ otherwise.

Example 17.3: M/M/s queue

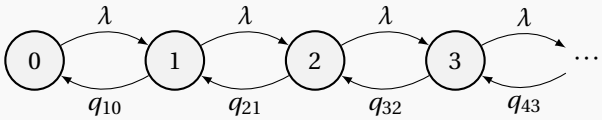
Customers arrive to a system with s servers according to a $\text{PP}(\lambda)$. If a server is available, the arrival immediately enters service. The service times are iid $\sim \text{Exp}(\mu)$. If no server is available, the arrival waits until one becomes available.

Let $(X_t)_{t \geq 0}$ denote the number of customers in the system at time $t \geq 0$. "In system" denotes customers that

are in queue or in service.

Your job is to model this as a CTMC, i.e. specify the jump rates.

The state space $\mathcal{S} = \{0, 1, 2, \ldots\}$. The only transitions that are possible are transitions one to the left or one to the right (the probability that two arrivals happen at the same time is 0 because it's a continuous process).



We have that $q_{n,n+1} = \lambda$; it's the arrivals of the Poisson process. We also have

$$q_{n,n-1} = \begin{cases} n\mu & 1 \le n \le s \\ s\mu & n > s \end{cases}.$$

The first case is where less than s servers are being used, and it's a race for which of those n finishes first; it's the minimum of n exponential RVs, and since they're all iid, this happens with a rate of $n\mu$. The second case is when more than s people are in the system, and so all s servers are being used.

Example 17.4: Birth-Death chain

Individuals give birth $\sim$ PP(λ) (all independently), and individuals have lifetimes of iid duration Exp(μ). Let X_t denote the number of individuals in population at time t.

The state space is $\mathcal{S} = \{0, 1, 2, \ldots\}$, and the transition rates are

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

This is because giving birth is like taking the minimum of n iid Exp(λ) RVs, and death is like taking the minimum of n iid Exp(μ) RVs.

Note that $q_n = n(\lambda + \mu)$, and $p_{n,n+1} = \frac{\lambda}{\lambda+\mu}$, $p_{n,n-1} = \frac{\mu}{\lambda+\mu}$; this means that the discrete time jump chain is also a birth-death chain.

17.1 Stationary Distributions

Definition 17.5: Rate Conservation Principle

A probability vector π is (with the exception of some weird cases) a stationary distribution for a CTMC with rate matrix $\mathbf{Q}$ if

$$\pi\mathbf{Q} = \vec{\mathbf{0}}.$$

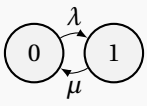
If we expand this out, it's equivalent to writing

$$\pi_j q_j = \sum_{i \in \mathcal{S}} \pi_i q_{ij} \quad \forall j \in \mathcal{S}.$$

The LHS is the rate at which transitions are made out of j, and the RHS is the rate at which transitions are made into j, all assuming that $\mathbb{P}(X_t = i) = \pi_i$.

Example 17.6

Consider a CTMC



Here, we have

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

Just like DTMCs, there is a classification of states in CTMCs:

- $i \leftrightarrow j \iff i \leftrightarrow j$ in jump chain, i.e. I can travel $i \to j$ and back. This means that classes in CTMCs are the same as those in the associated jump chain.
- State j is transient if, given $X_0 = j$, $(X_t)_{t\geq 0}$ re-enters state j finitely many times with probability one. State j is recurrent otherwise.
- For a recurrent state j, define the reentry time

  $$T_j = \min\{t \geq 0 : (X_t = j) \wedge (\exists s < t)(X_s \neq j)\}.$$

  That is, T_j is the first time we re-enter state j.
- State j is positive recurrent if $\mathbb{E}\left[T_j \mid X_0 = j\right] < \infty$.
- State j is null recurrent if $\mathbb{E}\left[T_j \mid X_0 = j\right] = \infty$.
- There is no concept of periodicity in CTMCs, because we can visit any accessible state in any (arbitrarily small) amount of time with positive probability.
- Transience/positive/null recurrence are class properties, just like in DTMCs

The Big Theorem characterizes the long-term behavior of CTMCs, just like we did for DTMCs (in Theorem 13.11).

Theorem 17.7: Big Theorem for CTMCs

For ease of writing, let's define $p_{ij}^t := \mathbb{P}\left(X_t = j \mid X_0 = i\right)$ (the transition probability for time t), and $m_j := \mathbb{E}\left[T_j \mid X_0 = j\right]$ (the mean recurrence time for state j).

For an irreducible CTMC, exactly one of the following is true:

1. All states are transient or null recurrent. No stationary distribution exists, and

   $$\lim_{t\to\infty} p_{ij}^t = 0 \quad \forall i, j \in \mathcal{S}.$$

2. All states are positive recurrent. There exists a unique stationary distribution, and satisfies

   $$\pi_j = \frac{1}{m_j q_j} = \lim_{t\to\infty} p_{ij}^t \quad \forall i, j \in \mathcal{S}.$$

Note: the stationary distribution in CTMC is not the same as the stationary distribution in the jump chain.

Generally speaking, $\tilde{\pi}_j \propto \pi_j q_j$, where $\tilde{\pi}_j$ is the stationary distribution in the jump chain. This requires $\sum q_i \pi_i < \infty$ or else weird things happen.

Example 17.8: M/M/∞ queue

Similar to the M/M/s queue, but we have infinite servers. That is, the transition rates

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

because arrivals $\sim$ PP(λ) and service times $\sim_{IID}$ Exp(μ).

If we solve $\pi\mathbf{Q} = \vec{\mathbf{0}}$, we find that

$$\pi_n = \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} e^{-\frac{\lambda}{\mu}}.$$

By the Big Theorem,

$$X_t \xrightarrow{d} \text{Pois}\left(\frac{\lambda}{\mu}\right).$$

In other words, the number of people in the system at time t converges in distribution to a Pois$\left(\frac{\lambda}{\mu}\right)$ distribution.

3/30/2021

Lecture 18

CTMC First Step Analysis, Uniformization, Random Graphs

18.1 First Step Analysis

This is exactly the same idea as for DTMCs. In fact, hitting probabilities (e.g. $\mathbb{P}$(reach A before B)) is exactly the same (we just look at the jump chain).

The only difference is when we consider time-dependent quantities (ex. expected hitting time).

If $A \subseteq \mathcal{S}$, define $T_A = \min\{t \geq 0 : X_t \in A\}$. Compute the expected hitting time, given that we start in state j, i.e. $\mathbb{E}\left[T_A \mid X_0 = j\right]$.

The strategy is the same as for DTMC, except we account for holding times. We define $t_i := \mathbb{E}[T_A \mid X_0 = i]$. That is, $t_i = 0 \; \forall i \in A$, and for $i \notin A$, we use first step analysis:

$$\underbrace{\mathbb{E}[T_A \mid X_0 = 0]}_{t_i} = \mathbb{E}[\text{time we hold in state } i] + \sum_{j \in \mathcal{S}} p_{ij}\, \mathbb{E}\left[T_A \mid X_0 = j\right].$$

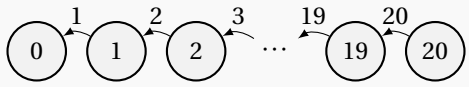
Here, we have p_{ij} as the transition probabilities in the jump chain. When we jump, it's as if time resets to 0 by the Markov Property. The system of first step equations that we obtain are

$$\begin{cases} t_i = 0 & \forall i \in A \\ t_i = \frac{1}{q_i} + \sum_{j \in \mathcal{S}} p_{ij} t_j & \forall i \notin A \end{cases}.$$

We can solve these equations to find t_i, $i \in \mathcal{S}$.

Example 18.1

If we have 20 lightbulbs with lifetimes $\sim_{IID}$ Exp(1) RVs, all on at time $t = 0$. How long does it take until they all burn out?



We can write down first step equations to compute

$$\begin{aligned} t_n &:= \mathbb{E}\left[T_{\{0\}} \mid X_0 = n\right] \\ t_1 &= 1 + t_0 = 1 \\ t_2 &= \frac{1}{2} + t_1 = \frac{1}{2} + 1 \\ t_3 &= \frac{1}{3} + t_2 = \frac{1}{3} + \frac{1}{2} + 1 \\ &\vdots \\ t_{20} &= 1 + \frac{1}{2} + \cdots + \frac{1}{20} \approx 3.6 \end{aligned}$$

18.2 Uniformization

"Uniformization" is essentially the simulation of a CTMC via a DTMC.

For context, consider a CTMC with transition rates $(q_i)_{i\in\mathcal{S}}$, and assume $\exists M > 0$ s.t. $q_i \le M \,\forall i \in \mathcal{S}$.

Let $\mathbf{P^t}$ denote the matrix of transition probabilities at time $t \ge 0$. That is,

$$[\mathbf{P^t}]_{ij} := \mathbb{P}\big(X_t = j \mid X_0 = i\big).$$

The Markov property gives $\mathbf{P^{s+t}} = \mathbf{P^s P^t}$, $\forall s, t \ge 0$.

We can show that $\mathbf{P^h} \approx \mathbf{I} + h\mathbf{Q} + o(h)$ for $h \approx 0$. So, we have

$$\begin{aligned} \mathbf{P^{t+h}} = \mathbf{P^t P^h} &= \mathbf{P^t}(\mathbf{I} + h\mathbf{Q} + o(h)) \\ \implies \frac{\mathbf{P}^{t+h} - \mathbf{P^t}}{h} &= \mathbf{P^t Q} + \frac{o(h)}{h} \\ \implies \frac{\partial}{\partial t}\mathbf{P}^t &= \mathbf{P}^t\mathbf{Q} \end{aligned}$$

This last equation is called the Kolmogorov Forward Equation. In particular, this differential equation has the unique solution

$$\mathbf{P}^t = e^{t\mathbf{Q}} = \sum_{k\ge 0} \frac{(t\mathbf{Q})^k}{k!} \quad \forall t \ge 0.$$

The question then becomes how do we compute $\mathbf{P}^t$ for large state spaces? Finding this matrix exponential is not easy; this is where uniformization comes in.

Definition 18.2: Uniformization

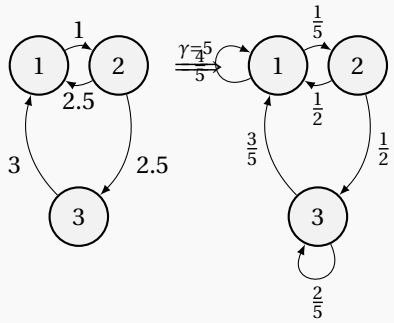
We take $\gamma \ge M$, which in turn $M \ge q_{ij}\,\forall i, j \in \mathcal{S}$.

Define a DTMC with transition probabilities

$$\begin{aligned} p_{ij} &= \frac{q_{ij}}{\gamma} \\ p_{ii} &= 1 - \frac{q_i}{\gamma} \end{aligned}$$

Note: these are not transition probabilities of the jump chain.

Example 18.3



If $\mathbf{P_u}$ is the transition matrix for a uniformized DTMC, then by construction,

$$\mathbf{P_u} = \mathbf{I} + \frac{1}{\gamma}\mathbf{Q}.$$

So observe that

$$\pi\mathbf{P_u} = \pi + \frac{1}{\gamma}\pi\mathbf{Q}.$$

That is, $\pi\mathbf{P_u} = \pi \iff \pi\mathbf{Q} = 0 \iff \pi$ is the stationary distribution for both the CTMC and the uniformized DTMC.

The point of uniformization is to compute (approximately) $\mathbf{P}^t$ by running the uniformized DTMC for some number of steps.

How does it work? We have that $\mathbf{P_u^n}$, the n-step transition probabilities for the uniformized DTMC, is equal to $\left(\mathbf{I} + \frac{1}{\gamma}\mathbf{Q}\right)^n \approx e^{\frac{n}{\gamma}\mathbf{Q}}$, because $1 + \varepsilon \approx e^{\varepsilon}$ for ε small.

So, to estimate $\mathbf{P}^t$, we run the uniformized chain for $n \approx \gamma t$ steps, because $\mathbf{P^t} = e^{t\mathbf{Q}} \approx e^{\frac{n}{\gamma}\mathbf{Q}} \approx \mathbf{P_u^n}$.

In summary, Euler schemes are discrete-time approximations to solving differential equations. Uniformization is an approach to compute continuous-time transition probabilities by simulating a DTMC.

18.3 Random Graphs

A lot of objects in EECS+ are modeled by graphs:

- social networks
- dependency structure in databases/programs
- tournaments
- epidemiology
- etc.

The simplest class of random graphs is given by the Erdös-Rënyi ensemble (i.e. the iid Bernoullis (coin flips) of the graph world).

Definition 18.4: Erdös–Rënyi random graphs

We fix $n \geq 1$ and $p \in [0,1]$. A random graph $\mathcal{G}(n,p)$ is an undirected graph on n vertices where each edge is placed independently with probability p.

Note here that we do not treat isomorphic graphs to be the same.

What types of question might we ask? If $n \to \infty$, how should p scale with n so that graph has a certain property $\mathcal{P}$ with high probability?

Like for capacity (in information theory), there's often a sharp "threshold" behavior.

Theorem 18.5: Monotone Graph Property Thresholds (Friedgut and Konlai, 1996)

Every "monotone" graph property has a sharp threshold t_n. (A monotone graph property is a property where adding more edges does not remove the property.)

That is,

$$p \gg t_n \implies \mathcal{G}(n,p) \text{ has } \mathcal{P} \text{ with high probability}$$
$$p \ll t_n \implies \mathcal{G}(n,p) \text{ doesn't have } \mathcal{P} \text{ with high property}$$

Example 18.6: Various property thresholds

1. If $p \ll \frac{1}{n^2}$ then there are no edges in $\mathcal{G}(n,p)$ with high probability (use Markov inequality).

   If $p \gg \frac{1}{n^2}$ then $\mathcal{G}(n,p)$ has edges with high probability.
2. If $p \gg \frac{1}{n}$ then $\exists$ cycle with high probability.

   If $p \ll \frac{1}{n}$ then $\nexists$ cycle with high probability.
3. If $p \ll \frac{1}{n}$ then the largest connected component is of size $\mathcal{O}(\log n)$ with high probability.

   If $p \gg \frac{1}{n}$ then the largest connected component is of size $\Theta(n)$.

   That is, $t_n = \frac{1}{n}$ is the threshold for the emergence of "giant component".

4/1/2021

Lecture 19

Connectivity Threshold, Inference

19.1 Connectivity Threshold

Recall that a graph is "connected" if there exists a path between any given pair of vertices.

Theorem 19.1: Connectivity Threshold (Erdös-Rënyi)

Fix $\lambda > 0$, and let $p_n = \lambda \frac{\log n}{n}$.

If $\lambda > 1$, then $\mathbb{P}\big(\mathcal{G}(n,p_n) \text{ is connected}\big) \to 1$.

If $\lambda < 1$, then $\mathbb{P}\big(\mathcal{G}(n,p_n) \text{ is connected}\big) \to 0$.

Proof. For the case $\lambda < 1$, we'll show something stronger than what we need:

$$\mathbb{P}\big(\mathcal{G}(n,p) \text{ contains isolated vertex}\big) \to 1.$$

This probability is always less $\leq \mathbb{P}\big(\mathcal{G}(n,p) \text{ disconnected}\big)$. Our strategy here is to let X be the number of isolated vertices. We'll use the lemma to show that $\mathbb{P}(X=0) \to 0$, which means that $\mathbb{P}(X \geq 1) \to 1$.

We can define indicators $I_i = \mathbf{1}_{\{\text{vertex } i \text{ is isolated}\}}$; this means that $X = \sum_{i=1}^n I_i$.

We know that $\mathbb{E}[X] = n\mathbb{E}[I_i] = n(1-p)^{n-1} = nq$, where $q = (1-p)^{n-1}$.

We can also compute

$$\begin{aligned}\text{Var}(X) &= \sum \text{Var}(I_i) + \sum_{i\neq j} \text{Cov}(I_i, I_j) \\ &= nq(1-q) + n(n-1)\frac{pq^2}{1-p}\end{aligned}$$

Using our lemma, we have that

$$\begin{aligned}\mathbb{P}(X=0) \leq \frac{nq(1-q)+n(n-1)\frac{pq^2}{1-p}}{n^2q^2} \\ &= \frac{1-q}{nq} + \frac{p}{1-p} \\ &\leq \frac{1}{nq} + \frac{p}{1-p}\end{aligned}$$

We know that p is vanishing, so the second term is going to 0. Hence, we just need to show that nq is going to infinity:

$$\log\big(nq\big) = \log\big(n(1-p)^{n-1}\big) = \log(n) + (n-1)\log\big(1-p\big).$$

With p vanishing, $\log\big(1-p\big) \approx p$, so this turns into

$$\log(n) - (n-1)p = \log n - (n-1)\lambda\frac{\log(n)}{n} \to (1-\lambda)\log(n) = \log\Big(n^{1-\lambda}\Big) \to \infty.$$

For the case of $\lambda > 1$, we have that

$$\begin{aligned}\mathbb{P}\big(\mathcal{G}(n,p) \text{ disconnected}\big) &= \mathbb{P}\left(\bigcup_{k=1}^{\frac{n}{2}} \{\exists \text{ set of } k \text{ disconnected vertices}\}\right) \\ &\leq \sum_{k=1}^{\frac{n}{2}} \mathbb{P}(\exists \text{ set of } k \text{ disconnected vertices}) \\ &\leq \sum_{k=1}^{\frac{n}{2}} \binom{n}{k} \mathbb{P}\big(\text{specific set of } k \text{ vertices disconnected from rest}\big) \quad \text{(union bound)} \\ &= \sum_{k=1}^{\frac{n}{2}} \binom{n}{k} (1-p)^{k(n-k)} \\ &\xrightarrow{\lambda>1} 0 \quad \text{(tedious)}\end{aligned}$$

In the last step, if we have two groups of $n-k$ and k vertices, we have $k(n-k)$ edges that aren't placed, i.e. $(1-p)^{k(n-k)}$. □

19.2 Statistical Inference

19.2.1 Hypothesis Testing

In hypothesis testing, we're trying to answer "What happened?", or "Did something happen?"

We start with some X, representing the state of nature; this takes values in $\{0,1,\cdots,M-1\}$, i.e. there are "M hypotheses to consider." This X is then transformed via $p_{Y|X}$, representing a model—these are some "likelihoods", perhaps pdfs or pmfs. This then outputs some observation Y.

X may or may not be a random variable, i.e. it may not be described by a known probability distribution. When X is a random variable with known distribution $\mathbb{P}(X = i) = \pi_i$ for $i = 0 \cdots M-1$, then we call π the "prior" (because it reflects our prior knowledge of the state of nature). This is what is called Bayesian Inference.

Example 19.2

Let $X = \{\text{Healthy}, \text{Covid}, \text{Flu}\}$, and $Y = \text{Symptoms observed}$.

An assumption we'll be making here is that we are always given the model $P_{Y|X}$.

If, for example, $\mathbb{P}(X = H) = 0.9$, $\mathbb{P}(X = F) = 0.03$, $\mathbb{P}(X = C) = 0.07$, then this is a prior π where $\pi(H) = 0.9$, $\pi(F) = 0.03$, and $\pi(C) = 0.07$. This reflects our prior knowledge of the state of nature, (eg. prevalence of flu/covid in the general population).

By Bayes rule, if we observe $\{Y = y\}$, then the "a posteriori" probability of $\{X = x\}$ is given by:

$$\mathbb{P}\big(X = x \mid Y = y\big) = \frac{P_{Y|X}(y \mid x)\pi(x)}{\sum_{\tilde{x}} P_{Y|X}(y \mid \tilde{x})\pi(\tilde{x})} \propto P_{Y|X}(y \mid x)\pi(x).$$

We can think of this as an update of the prior, given our observations. Note here that the denominator does not depend on x, just y.

So, this motivates the Maximum A Posteriori (MAP) estimate. In other words, the most likely x after observing $\{Y = y\}$ is given by

$$\hat{X}_{MAP}(y) = \operatorname*{argmax}_x P_{X|Y}(x \mid y) = \operatorname*{argmax}_x P_{Y|X}(y \mid x)\pi(x).$$

The MAP estimate depends on the likelihoods and the prior.

As an exercise, one can show that for any other $\hat{X}(y)$,

$$\mathbb{P}\big(\hat{X}(y) \neq x \mid Y = y\big) \geq \mathbb{P}\big(\hat{X}_{MAP}(y) \neq x \mid Y = y\big).$$

What if we don't have a prior? One strategy is to assume that π is uniform over all x. In this case, the MAP estimate reduces to maximizing the likelihood of the observation over the hypotheses. This gives rise to the Maximum Likelihood (ML) estimate:

$$\hat{X}_{ML}(y) = \operatorname*{argmax}_x P_{Y|X}(y \mid x).$$

Note that the notes use the notation of $\text{MAP}(X \mid Y = y)$ and $\text{MLE}(X \mid Y = y)$, but it's clunky so we aren't going to use it here.

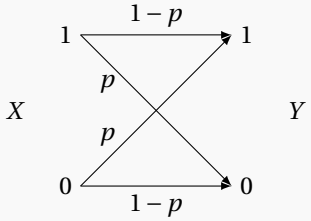
4/8/2021

Lecture 20

Binary Hypothesis Testing

Example 20.1

Consider a $BSC(p)$, with X as its input and Y as its output.



We can calculate the ML estimate. If $p < \frac{1}{2}$, then

$$\hat{X}_{ML}(y) = \operatorname*{argmax}_{x\in\{0,1\}} P_{Y|X}(y \mid x) = y.$$

If $p > \frac{1}{2}$, then the opposite is true (i.e. $\hat{X}_{ML}(y) = 1 - y$).

MAP is a little bit more complicated. Recall that we want to find the x that maximizes $P_{Y|X}(y \mid x)\pi(x)$.

For $y = 0$, we have $P_{Y|X}(0 \mid x)\pi(x) = \begin{cases} (1-p)\pi_0 & x = 0 \\ p(1-\pi_0) & x = 1 \end{cases}$. Here, π_0 is the prior probability that $X = 0$.

This means that $\hat{X}_{MAP}(0) = \begin{cases} 0 & p < \pi_0 \\ 1 & p \geq \pi_0 \end{cases}$.

Why is this the case? $x = 0$ gives the maximum when $(1-p)\pi_0 > p(1-\pi_0)$, or when $\pi_0 > p$. A similar simplification can be done for $x = 1$, or it can be reasoned as it's the only other option.

Repeating this for $y = 1$, we see that $P_{Y|X}(1 \mid x) = \begin{cases} p\pi_0 & x = 0 \\ (1-p)(1-\pi_0) & x = 1 \end{cases}$.

This means that $\hat{X}_{MAP}(1) = \begin{cases} 0 & 1-p < \pi_0 \\ 1 & 1-p \geq \pi_0 \end{cases}$.

Definition 20.2: Likelihood Ratio

In a problem of binary hypothesis testing (i.e. $M = 2$; there are only two hypotheses), we can define the *likelihood ratio*:

$$L(y) := \frac{P_{Y|X}(y \mid 1)}{P_{Y|X}(y \mid 0)}.$$

This means that we can reformulate the previous example (and any binary hypothesis test) as:

$$\hat{X}_{ML}(y) = \begin{cases} 1 & L(y) \geq 1 \\ 0 & L(y) < 1 \end{cases}$$

$$\hat{X}_{MAP}(y) = \begin{cases} 1 & L(y) \geq \frac{\pi_0}{\pi_1} \\ 0 & L(y) < \frac{\pi_0}{\pi_1} \end{cases}$$

This means that both $\hat{X}_{ML}$ and $\hat{X}_{MAP}$ can be expressed as a "threshold test" where we just threshold likelihood ratio evaluated for the observation!

Definition 20.3: Threshold test

Threshold tests are decision rules of the form

$$\hat{X}(y) = \begin{cases} 1 & L(y) > \lambda \\ 0 & L(y) < \lambda \\ \text{Bernoulli}(\gamma) & L(y) = \lambda \end{cases}.$$

The last case states that in case of a tie, we break them randomly.

Deriving the above two quantities is as follows:

Proof. We have that $\hat{X}_{MAP}(y) = \operatorname{argmax}_x P_{Y|X}\pi(x)$. This is equivalent to saying that we choose $x = 1$ if:

$$\begin{aligned} P_{Y|1}(y \mid 1)\pi(1) &\geq P_{Y|0}(y \mid 0)\pi(0) \\ \frac{P_{Y|1}(y \mid 1)}{P_{Y|0}(y \mid 0)} &\geq \frac{\pi(0)}{\pi(1)} \\ L(y) &\geq \frac{\pi(0)}{\pi(1)} \end{aligned}$$

The ML estimate $\hat{X}_{ML}$ can be derived in the same way, just with $\pi(0) = \pi(1)$, meaning $\frac{\pi(0)}{\pi(1)} = 1$. □

We've talked about this in the discrete case, but continuous observations work in the same way.

Example 20.4

Suppose we have $X \in \{0,1\}$, and $Y = X + Z$ where $Z \sim \mathcal{N}(0,\sigma^2)$ independent of X.

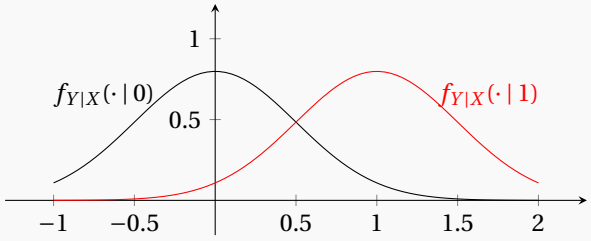
The likelihoods can be calculated as

$$\begin{aligned} f_{Y|X}(y \mid 0) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right) \\ f_{Y|X}(y \mid 1) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-1)^2}{2\sigma^2}\right) \end{aligned}$$

The likelihood ratio is then (simplification omitted)

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

This scenario can be depicted as follows:



We then have

$$\hat{X}_{MAP}(y) = \begin{cases} 1 & L(y) \geq \frac{\pi_0}{\pi_1} \iff y \geq \frac{1}{2} + \sigma^2 \log\left(\frac{\pi_0}{\pi_1}\right) \\ 0 & L(y) < \frac{\pi_0}{\pi_1} \iff y < \frac{1}{2} + \sigma^2 \log\left(\frac{\pi_0}{\pi_1}\right) \end{cases}$$

$$\hat{X}_{ML}(y) = \begin{cases} 1 & L(y) \geq 1 \iff y \geq \frac{1}{2} \\ 0 & L(y) < 1 \iff y < \frac{1}{2} \end{cases}$$

20.1 Binary Hypothesis Testing

The previous examples are instances of "binary hypothesis testing"—that is, $X \in \{0,1\}$. We have two hypotheses to discriminate between, given our observation y:

$$\begin{aligned} H_0 &: Y \sim P_{Y|X=0} && \text{(null hypothesis)} \\ H_1 &: Y \sim P_{Y|X=1} && \text{(alternate hypothesis)} \end{aligned}$$

In other words, the null hypothesis is under the assumption that $X = 0$, and the alternate hypothesis is under the assumption that $X = 1$.

In the end, we want a decision rule (a "test"): $\hat{X} : y \to \{0,1\}$.

Definition 20.5: Type I and II Error Probability

With any test, there are two fundamental types of error:

- **Type I Error Probability** (false positive probability):

$$\mathbb{P}\big(\hat{X}(Y) = 1 \mid X = 0\big).$$

- **Type II Error Probability** (true negative probability):

$$\mathbb{P}\big(\hat{X}(Y) = 0 \mid X = 1\big).$$

Our goal is to choose a test that minimizes the Type II error probability subject to a constraint on Type I error probability.

In other words, for $\beta \geq 0$, we want to find

$$\hat{X}^*_\beta = \operatorname*{argmin}_{\hat{X}:\mathbb{P}(\hat{X}(Y)=1|X=0)<\beta} \mathbb{P}\big(\hat{X}(Y) = 0 \mid X = 1\big).$$

Theorem 20.6: Neyman–Pearson Lemma

Given $\beta \in [0,1]$, the optimal decision rule is a (randomized) threshold test of the following form:

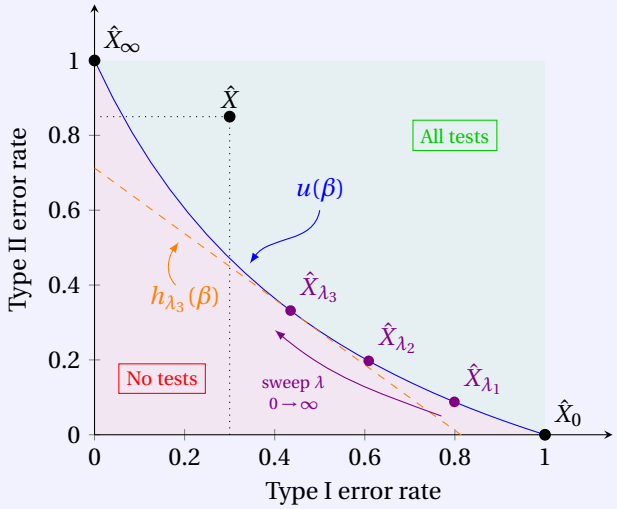
$$\hat{X}_\beta(y) = \begin{cases} 1 & L(y) > \lambda \\ 0 & L(y) < \lambda \\ \text{Bernoulli}(\gamma) & L(y) = \lambda \end{cases},$$

where λ, γ are chosen so that $\mathbb{P}\big(\hat{X}(Y) = 1 \mid X = 0\big) = \beta$ (i.e. the type I error probability = β).

Proof from lecture 21:

Proof. There are two things we care about for a test $\hat{X} : Y \to \hat{X}(Y) \in \{0,1\}$: the Type I and Type II error rates.

We can plot the Type I and Type II error rates. Here, we let the threshold test with threshold λ be denoted by $\hat{X}_\lambda$.



The function $u(\beta)$ is the "error curve". We can express this function as follows:

$$u(\beta) := \max_{\lambda \geq 0} \{\mathbb{P}\big(\hat{X}_\lambda(Y) = 0 \mid X = 1\big) + \lambda\big(\mathbb{P}\big(\hat{X}_\lambda(Y) \mid X = 0\big) - \beta\big)\}.$$

Although this looks complicated, note that the inside of the max is $h_\lambda(\beta)$, which is just an affine function of β for a fixed λ. This is because when we fix λ, all the probabilities are functions, and the only thing that varies is β.

Our goal is to show that all threshold tests lie precisely on the error curve, and to show that all other tests lie above the error curve (i.e. no other test lies below).

First, observe that for a fixed λ_0,

$$u(\mathbb{P}\big(\hat{X}_{\lambda_0}(Y) = 1 \mid X = 0\big)) \geq \underbrace{\mathbb{P}\big(\hat{X}_{\lambda_0}(Y) = 0 \mid X = 1\big)}_{\text{Type II error probability for } \hat{X}_{\lambda_0}}$$

We get this bound because plugging in the LHS into the definition of $u(\beta)$ makes the second term in the maximum 0—choosing $\lambda = \lambda_0$ arbitrarily gives us this bound on the test for λ_0.

What this is saying is that $\hat{X}_{\lambda_0}$ lies on or below the error curve; we plug in the Type I error probability into u, and we've made a bound for the Type II error probability. Since we picked this λ_0 arbitrarily, all threshold tests therefore must lie on or below the error curve as well.

Now, we will show that *all* tests lie above the error curve.

Suppose we fix $\lambda \in [0, \infty)$. Note that there is no prior on X here—hence, we'll impose an artificial prior on X with $\frac{\pi_0}{\pi_1} = \lambda$.

In this case, $\hat{X}_{MAP}(Y) = \hat{X}_{\frac{\pi_0}{\pi_1}}(Y) = \hat{X}_\lambda(Y)$. The MAP test has the crucial property that it minimizes the probability of error:

$$\mathbb{P}\big(\hat{X}_{MAP}(Y) \neq X\big) \leq \mathbb{P}\big(\hat{X}(Y) \neq X\big) \text{ for any test } \hat{X}.$$

(We can't just always use this because the MAP test requires a prior.)

This means that if we write out the probability of error (using the law of total probability),

$$\pi_0 \mathbb{P}\big(\hat{X}(Y) = 1 \mid X = 0\big) + \pi_1 \mathbb{P}\big(\hat{X}(Y) = 0 \mid X = 1\big) \geq \pi_0 \mathbb{P}\big(\hat{X}_\lambda(Y) = 1 \mid X = 0\big) + \pi_1 \mathbb{P}\big(\hat{X}(Y) = 0 \mid X = 1\big) a$$

Dividing by π_1, we have

$$\mathbb{P}\big(\hat{X}(Y) = 0 \mid X = 1\big) + \lambda \mathbb{P}\big(\hat{X}(Y) = 1 \mid X = 0\big) \geq \mathbb{P}\big(\hat{X}_\lambda(Y) = 0 \mid X = 1\big) + \lambda \mathbb{P}\big(\hat{X}_\lambda = 1 \mid X = 0\big)$$

Moving $\lambda \mathbb{P}\big(\hat{X}(Y) = 1 \mid X = 0\big)$ to the RHS and collecting the λs, we have

$$\mathbb{P}\big(\hat{X}(Y) = 0 \mid X = 1\big) \geq \mathbb{P}\big(\hat{X}_\lambda(Y) = 0 \mid X = 1\big) + \lambda\big(\mathbb{P}\big(\hat{X}_\lambda = 1 \mid X = 0\big) - \mathbb{P}\big(\hat{X}(Y) = 1 \mid X = 0\big)\big)$$

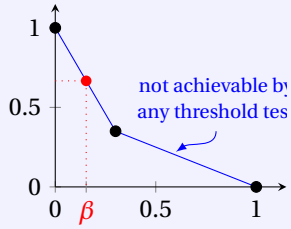
Note that λ was arbitrary. If we maximize the RHS over all λ, this turns into the u function:

$$\underbrace{\mathbb{P}\big(\hat{X}(Y) = 0 \mid X = 1\big)}_{\text{Type II error probability}} \geq u(\underbrace{\mathbb{P}\big(\hat{X}(Y) = 1 \mid X = 0\big)}_{\text{Type I error probability}}).$$

This means that $\hat{X}$ lies on or above the error curve u.

Since we've proven that all tests lie on or above the error curve, but threshold tests lie on or below the error curve, we've shown that threshold tests must lie exactly on the error curve.

Where does the randomization enter the picture? Threshold tests don't always continuously sweep out a curve u. For example, we could have the following, where all threshold tests lie on one of three discrete points:



How would we achieve this red point? We'd just flip a coin between two threshold tests—the type I and type II errors will be a weighted average of the two tests, depending on the bias of the coin (by law of total probability). □

$\hat{X}^*_\beta$ is called the "Neyman–Pearson Rule". It is the most powerful test (minimizes type II error) subject to the constraint that $\mathbb{P}(\text{Type I error} \le \beta)$.

4/13/2021

Lecture 21

Estimation

Example 21.1

If we go back to the Gaussian example, we had $X = 0 \implies Y \sim \mathcal{N}(0, \sigma^2)$ and $X = 1 \implies Y \sim \mathcal{N}(1, \sigma^2)$.

The likelihood ratio was calculated to be $L(y) = \exp\left(\frac{y}{\sigma^2} - \frac{1}{2\sigma^2}\right)$.

Let's say I want the Type I error probability to be $\le \beta$. How do we choose the optimal test (i.e. the one that minimizes the type II error rate)?

Neyman–Pearson tells us we should only consider threshold tests, and we just need to choose λ, γ appropriately to meet the type I error constraint.

Note that $\mathbb{P}(L(Y) = \lambda) = 0$ for any λ because Y is a continuous RV. Hence, no randomization is needed here.

Now the question is how we choose λ? We can start by writing out the Type I error constraint:

$$\begin{aligned}
\beta &= \overbrace{\mathbb{P}\left(\hat{X}(Y) = 1 \mid X = 0\right)}^{\text{Type I error probability}} \\
&= \mathbb{P}(\underbrace{Y \ge \frac{1}{2} + \sigma^2 \log \lambda}_{\{L(Y) \ge \lambda\}} \mid X = 0) \\
&= \mathbb{P}\left(\frac{Y}{\sigma} \ge \frac{1}{2\sigma} + \sigma \log \lambda \mid X = 0\right) \\
&= 1 - \Phi\left(\frac{1}{2\sigma} + \sigma \log \lambda\right)
\end{aligned}$$

The last step uses the fact that Y/σ would normalize the normal distribution $Y \sim \mathcal{N}(0, \sigma^2)$ (as we know $X = 0$), turning it into a standard normal. We'd then just solve for λ here in terms of β and σ.

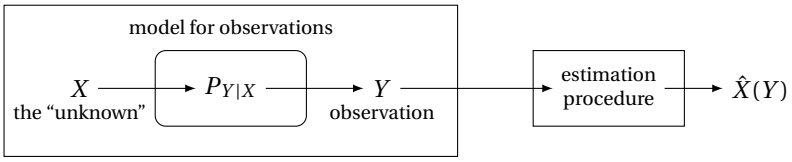
21.1 Estimation

Hypothesis testing tries to discriminate between two (or more) discrete hypotheses. Estimation is another inference problem, but now we try to guess the numerical value of some unknown quantity.

For example, given measurements by GPS, what is my latitude/longitude? The measurements are noisy, so we need to infer the position. There's also a notion of "closeness"; getting a position within 1 meter is better than 100 meters. Or, given the history of a stock, what will be the value tomorrow?

These types of problems are common in ML, communications, signal processing, finance, etc.

The general setup:



The "model", assumed to be given.
(i.e. we know P_{XY})

Our goal is to choose $\hat{X}$ to make the mean squared error (MSE) as small as possible.

Definition 21.2: Mean Squared Error

The *mean squared error* is $\mathbb{E}\left[(X-\hat{X}(Y))^2\right]$, abbreviated as MSE.

A fact (from HW) is that

$$\mathbb{E}[X \mid Y] = \operatorname*{argmin}_{\hat{X}} \mathbb{E}\left[(X-\hat{X}(Y))^2\right].$$

In other words, $\mathbb{E}[X \mid Y]$ minimizes the mean squared error.

So, the MSE estimation problem is totally solved in this sense. But, this is not practical in many cases; $\mathbb{E}[X \mid Y]$ is hard to compute, even if we know P_{XY} exactly (it usually involves integration). Further, it's often the case that we don't know P_{XY} exactly, but just have some reasonable model of it.

The workaround is to focus on *linear estimation*. That is, we focus on estimators that are linear functions of our observations (in general, $\mathbb{E}[X \mid Y]$ is nonlinear). We'd minimize the MSE over all "linear estimators" of the form

$$\hat{X}(Y) = a + \sum_{i=1}^{n} b_i Y_i, \text{ where } Y = (Y_1, \cdots, Y_n) = \text{vector of observations.}$$

This problem is called linear least squares estimation. The best linear estimator is called the *linear least squares estimate* (LLSE), denoted by $\mathbb{L}[X \mid Y]$. This notation reminds us of $\mathbb{E}[X \mid Y]$, but $\mathbb{L}$ for *linear*.

LLS estimation is just linear algebra disguised as probability.

4/15/2021

Lecture 22

Linear Estimation

22.1 Linear Estimation

Definition 22.1: Linear Least Squares Estimate

The best linear estimator is called the *linear least squares estimate* (LLSE), denoted by $\mathbb{L}[X \mid Y]$:

$$\mathbb{L}[X \mid Y] = \underset{\text{linear } \hat{X}}{\operatorname{argmin}} \mathbb{E}\left[\left|X - \hat{X}(Y)\right|^2\right],$$

where linear estimators $\hat{X}(Y)$ are of the form

$$\hat{X}(Y) = a + \sum_{i=1}^{n} b_i Y_i,\ a, b_1, \cdots, b_n \in \mathbb{R}.$$

A question we need to answer is how we'd solve for $\mathbb{L}[X \mid Y]$.

We want to solve for:

$$\min_{a, b_1, \ldots, b_n} \mathbb{E}\left[\left|X - (a + \sum b_i Y_i)\right|^2\right].$$

22.1.1 Calculus Approach

A first approach could be with calculus. If we expand out this square, we'd get

$$\begin{aligned} J(a, b_1, \cdots, b_n) &:= \mathbb{E}\left[\left|X - (a + \sum b_i Y_i)\right|^2\right] \\ &= \mathbb{E}[X]^2 - 2a\mathbb{E}[X] - 2\sum b_i \mathbb{E}[XY_i] + a^2 + 2a\sum b_i \mathbb{E}[Y_i] + \sum b_i^2 \mathbb{E}\left[Y_i^2\right] + \sum_{i \neq j} \mathbb{E}\left[Y_i Y_j\right] \end{aligned}$$

We'd then take the partial derivative with respect to a and the b_i's, and set to zero:

$$\begin{aligned} \frac{\partial}{\partial a} J = 0 \implies & \qquad a = \mathbb{E}[X] - \sum_{i=1}^{n} b_i \mathbb{E}[Y_i] \\ \frac{\partial}{\partial b_i} J = 0 \implies & \qquad \mathbb{E}[XY_i] = a\mathbb{E}[Y_i] + b_i \mathbb{E}\left[Y_i^2\right] + \sum_{i \neq j} b_j \mathbb{E}\left[Y_i Y_j\right] \end{aligned}$$

This is a system of linear equations we can solve for $a, b_1, \ldots, b_n$.

Let's make life easy by assuming $\mathbb{E}[X] = \mathbb{E}[Y_i] = 0$ for all i—this eliminates a bunch of terms. We'd see that

$$\begin{aligned} a &= 0 \\ \mathbb{E}[XY_i] &= \sum_{j=1}^{n} b_j \mathbb{E}\left[Y_i Y_j\right] \end{aligned}$$

Now, let us define a few terms. Let $\Sigma_{\mathbf{XY}} = \mathbb{E}\left[(X - \mu_X)(\vec{\boldsymbol{Y}} - \mu_{\vec{\boldsymbol{Y}}})^T\right]$ and $\Sigma_{\mathbf{Y}} = \mathbb{E}\left[(\vec{\boldsymbol{Y}} - \mu_{\vec{\boldsymbol{Y}}})(\vec{\boldsymbol{Y}} - \mu_{\vec{\boldsymbol{Y}}})^T\right]$. Here, we have $\vec{\boldsymbol{Y}} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$, $\mu_{\vec{\boldsymbol{Y}}} = \mathbb{E}\left[\vec{\boldsymbol{Y}}\right]$, and $\mu_X = \mathbb{E}[X]$. The matrix $\Sigma_{\mathbf{XY}}$ is often called the "cross-covariance matrix" (here it's a row vector because X is a scalar) because each element corresponds to the covariance between X and Y_i. Similarly $\Sigma_{\mathbf{Y}}$ is often called the "covariance matrix", because each element corresponds to the covariance between Y_i and Y_j.

The second equation previously can be rewritten as $\Sigma_{\mathbf{XY}} = \vec{\boldsymbol{b}}^T \Sigma_{\mathbf{Y}}$, where $\vec{\boldsymbol{b}} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$.

This tells us that $\vec{\boldsymbol{b}}^T = \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1}$. In particular, if $\mathbb{E}[X] = \mathbb{E}[Y_i] = 0$ for all i, then $\mathbb{L}[X \mid Y] = \vec{\boldsymbol{b}}^T \vec{\boldsymbol{Y}} = \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \vec{\boldsymbol{Y}}$.

If we don't have zero-mean, we'd just add the means back in to get:

$$\mathbb{L}[X \mid Y] = \mu_X + \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} (\vec{\boldsymbol{Y}} - \mu_{\vec{\boldsymbol{Y}}}).$$

Note that if $\vec{\boldsymbol{X}}$ is a vector—that is, $\vec{\boldsymbol{X}} = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}$—then the linear estimation problem of the form

$$\min_{\text{linear } \hat{\vec{\boldsymbol{X}}}} \mathbb{E}\left[\left|X - \hat{\vec{\boldsymbol{X}}}(Y)\right|^2\right] = \sum_{i=1}^{k} \min_{\text{linear } \hat{X}_i} \mathbb{E}\left[\left|X_i - \hat{X}_i(Y)\right|^2\right].$$

So, we can always assume X is a scalar, because vector problems decompose into scalar problems.

Nevertheless, the expression we calculated before for $\mathbb{L}[X \mid Y]$ remains valid for a vector-valued X.

Observe that $\mathbb{L}[X \mid Y]$ only depends on the first and second order statistics of X and Y (i.e. means and covariances). In practice, this is good, because we rarely know the joint distribution of X and Y completely, but we can estimate the first and second order statistics from data.

Also observe that linear estimation doesn't always mean linear:

Example 22.2

Let X, Y be random variables; how would we compute the best quadratic estimator of the form

$$\hat{X}(Y) = a + b_1 Y + b_2 Y^2?$$

We'd just compute this as

$$\hat{X}_Q(Y) = \mu_X + \Sigma_{\mathbf{X}\tilde{\mathbf{Y}}} \Sigma_{\tilde{\mathbf{Y}}}^{-1} (\tilde{Y} - \mu_{\tilde{Y}}),$$

where $\tilde{Y} = \begin{bmatrix} Y \\ Y^2 \end{bmatrix}$.

As such, linear estimation doesn't mean that the data is linear; we're just restricting the class of possible estimators to just linear functions of whatever we're trying to compute.

22.1.2 Connection to linear regression

Consider a simple observation model of the form $\vec{\boldsymbol{Y}} = \mathbf{A}\vec{\boldsymbol{X}} + \vec{\boldsymbol{Z}}$, where $\mathbf{A} \in \mathbb{R}^{n \times k}$, and as such $\vec{\boldsymbol{Y}}$ is an n-vector and $\vec{\boldsymbol{X}}$ is a k-vector. Further let $\Sigma_{\mathbf{X}} = \sigma_{\vec{\boldsymbol{X}}}^2 \mathbf{I}$, $\Sigma_{\mathbf{Z}} = \sigma_{\vec{\boldsymbol{Z}}}^2 \mathbf{I}$, where $\vec{\boldsymbol{X}}$ and $\vec{\boldsymbol{Z}}$ are uncorrelated (i.e. elements of $\vec{\boldsymbol{X}}$ are also uncorrelated with one another, and similarly for elements of $\vec{\boldsymbol{Z}}$; $\Sigma_{\mathbf{X}}$ and $\Sigma_{\mathbf{Z}}$ are both diagonal matrices with the variances along the main diagonal).

We'll assume everything is zero-mean for simplicity.

The best linear estimator is

$$\mathbb{L}\left[\vec{\boldsymbol{X}} \mid \vec{\boldsymbol{Y}}\right] = \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \vec{\boldsymbol{Y}}.$$

Note that

$$\begin{aligned}\Sigma_{\mathbf{XY}} &= \mathbb{E}\left[\vec{\boldsymbol{X}}\vec{\boldsymbol{Y}}^T\right] \\ &= \mathbb{E}\left[\vec{\boldsymbol{X}}(\vec{\boldsymbol{X}}^T\vec{\boldsymbol{A}}^T + \vec{\boldsymbol{Z}}^T)\right] \\ &= \sigma^2_{\vec{\boldsymbol{X}}}\mathbf{A}^T \\ \Sigma_{\mathbf{Y}} &= \mathbb{E}\left[\vec{\boldsymbol{Y}}\vec{\boldsymbol{Y}}^T\right] \\ &= \mathbb{E}\left[(\mathbf{A}\vec{\boldsymbol{X}} + \vec{\boldsymbol{Z}})(\mathbf{A}\vec{\boldsymbol{X}} + \vec{\boldsymbol{Z}})^T\right] \\ &= \sigma^2_{\vec{\boldsymbol{X}}}\mathbf{A}\mathbf{A}^T + \sigma^2_{\vec{\boldsymbol{Z}}}\mathbf{I}\end{aligned}$$

This is a result of a "Bayesian setting", where we assume we know $\sigma^2_{\vec{\boldsymbol{X}}}$. If we didn't know $\sigma^2_{\vec{\boldsymbol{X}}}$, the best we can do is assume $\sigma^2_{\vec{\boldsymbol{X}}} = +\infty$.

In this case, we have

$$\mathbb{L}_{minimax}[\vec{\boldsymbol{X}} \mid \vec{\boldsymbol{Y}}] = \lim_{\sigma^2_{\vec{X}} \to \infty} \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T + \frac{\sigma^2_{\vec{\boldsymbol{Z}}}}{\sigma^2_{\vec{\boldsymbol{X}}}}\mathbf{I}\right)^{-1}\vec{\boldsymbol{Y}}.$$

This inverse is the limit definition of the left inverse of $\mathbf{A}$, assuming full column rank; this means that this expression turns into

$$\mathbb{L}_{minimax}[\vec{\boldsymbol{X}} \mid \vec{\boldsymbol{Y}}] = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\vec{\boldsymbol{Y}}.$$

Recall the linear regression problem, where we try to minimize $\min_x \left\|\mathbf{A}\vec{\boldsymbol{x}} - \vec{\boldsymbol{y}}\right\|^2$. We saw from 16A that the least-squares solution is

$$\hat{\vec{\boldsymbol{x}}}_{LS} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\vec{\boldsymbol{y}}.$$

This is exactly what we arrived at previously—the moral of the story here is that linear regression can be considered a special case of linear estimation (which is a non-Bayesian, linear observation model).

22.2 Geometry of Linear Estimation

Let $\mathcal{V}$ be a vector space over a real scalar field. Let $\langle\cdot,\cdot\rangle$ be an inner product on $\mathcal{V}$. That is,

1. (Symmetry) $\langle\vec{\boldsymbol{u}},\vec{\boldsymbol{v}}\rangle = \langle\vec{\boldsymbol{v}},\vec{\boldsymbol{u}}\rangle$ for $\vec{\boldsymbol{u}},\vec{\boldsymbol{v}} \in \mathcal{V}$
2. (Linearity) $\langle a\vec{\boldsymbol{u}} + b\vec{\boldsymbol{v}},\vec{\boldsymbol{w}}\rangle = a\langle\vec{\boldsymbol{u}},\vec{\boldsymbol{w}}\rangle + b\langle\vec{\boldsymbol{v}},\vec{\boldsymbol{w}}\rangle$ for $a,b \in \mathbb{Z}$ and $\vec{\boldsymbol{u}},\vec{\boldsymbol{v}},\vec{\boldsymbol{w}} \in \mathcal{V}$
3. $\langle\vec{\boldsymbol{u}},\vec{\boldsymbol{u}}\rangle \geq 0$ for all $\vec{\boldsymbol{u}} \in \mathcal{V}$, and $\langle\vec{\boldsymbol{u}},\vec{\boldsymbol{u}}\rangle = 0 \iff \vec{\boldsymbol{u}} = \vec{\boldsymbol{0}}$.

$\mathcal{V}$ is called a (real) inner product space. It is a normed vector space, with norm

$$\|\vec{\boldsymbol{v}}\| := \sqrt{\langle\vec{\boldsymbol{v}},\vec{\boldsymbol{v}}\rangle}.$$

Norms satisfy

1. $\langle a\vec{\boldsymbol{v}}\rangle = |a|\|\vec{\boldsymbol{v}}\|$ for $a \in \mathbb{R}$, $\vec{\boldsymbol{v}} \in \mathcal{V}$
2. $\|\vec{\boldsymbol{v}}\| \geq 0$ and $\|v\| = 0 \iff \vec{\boldsymbol{v}} = \vec{\boldsymbol{0}}$
3. (Triangle inequality) $\|\vec{\boldsymbol{u}} + \vec{\boldsymbol{v}}\| \leq \|\vec{\boldsymbol{u}}\| + \|\vec{\boldsymbol{v}}\|$

$\mathcal{V}$ is called a "Hilbert Space" if it is "complete" with respect to its norm $\|\cdot\|$. Completeness just means that we can take limits without popping out of the space.

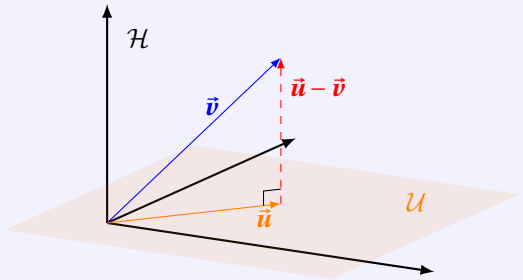
For example, $C_b(\mathbb{R})$ = continuous bounded functions on $\mathbb{R}$ is complete with respect to the norm $\|f\| = \max_x |f(x)|$, but is not complete with respect to the norm $\|f\| = \int |f|^2 \, \mathrm{d}x$.

Hilbert spaces enjoy a notion of geometry compatible with our intuition.

Theorem 22.3: Hilbert Projection Theorem

Let $\mathcal{H}$ be a Hilbert space, and $\mathcal{U} \subseteq \mathcal{H}$ be a closed subspace. For each $\vec{v} \in \mathcal{H}$, there is a unique closest point $\vec{u} \in \mathcal{U}$ to $\vec{v}$, i.e. $\operatorname{argmin}_{u \in \mathcal{U}}$ exists and is unique.

Moreover, $\vec{u} \in \mathcal{U}$ is the closest point to $\vec{v} \in \mathcal{H}$ if and only if $\langle \vec{u} - \vec{v}, \vec{u}' \rangle = 0$, $\forall \vec{u}' \in \mathcal{U}$ (i.e. orthogonal).



Note that we also have the Pythagorean Theorem:

$$\|\vec{u}\|^2 + \|\vec{u} - \vec{v}\|^2 = \|\vec{v}\|^2.$$

A short proof is as follows:

$$\begin{aligned}\langle \vec{u}, \vec{u} \rangle + \langle \vec{u} - \vec{v}, \vec{u} - \vec{v} \rangle &= \langle \vec{u}, \vec{u} \rangle + \langle \vec{u} - \vec{v}, \vec{u} \rangle - \langle \vec{u} - \vec{v}, \vec{v} \rangle \\ &= 2\langle \vec{u}, \vec{u} - \vec{v} \rangle^{0} + \langle \vec{v}, \vec{v} \rangle \\ &= \|\vec{v}\|^2\end{aligned}$$

Definition 22.4: Hilbert Space of Random Variables

Let $(\Omega, \mathcal{F}, P)$ be a probability space. The collection of random variables X with finite second moments (i.e. $\mathbb{E}[X^2] < \infty$ form a Hilbert space with respect to the inner product

$$\langle X, Y \rangle := \mathbb{E}[XY].$$

In this notation, $\|X\|^2 = \mathbb{E}[X^2]$.

22.3 Connection to Linear Estimation

Theorem 22.5: Orthogonality Principle

For RVs $Y_1, \cdots, Y_n$ with finite second moments, the following equations uniquely characterize $\mathbb{L}[X \mid Y]$:

$$\mathbb{E}[\mathbb{L}[X \mid Y]] = \mathbb{E}[X]$$
$$\mathbb{E}[\mathbb{L}[X \mid Y]\, Y_i] = \mathbb{E}[XY_i]$$

Proof. For RVs $Y_1, \cdots, Y_n$ with finite second moments, the space of RVs

$$\mathcal{U} = \{a + \sum b_i Y_i : a, b_1, \cdots, b_n \in \mathbb{R}\}$$

is a closed subspace of the Hilbert space of RVs.

By the Hilbert Projection Theorem, $\mathbb{L}[X \mid Y] = \operatorname{argmin}_{u \in \mathcal{U}} \|X - u\|^2$ exists and is unique. We can rewrite this (by definition of the norm) as $\mathbb{L}[X \mid Y] = \operatorname{argmin}_{\text{linear } \hat{X}} \mathbb{E}\left[(X - \hat{X}(Y))\right]^2$.

Moreover, it is characterized by equations:

$$\langle \mathbb{L}[X \mid Y] - X, u\rangle = \mathbb{E}[(\mathbb{L}[X \mid Y] - X)u] = 0 \qquad \forall u \in \mathcal{U}$$
$$\mathbb{E}\left[(\mathbb{L}[X \mid Y] - X)\left(a + \sum b_i Y_i\right)\right] = 0 \qquad \forall a, b_i$$

With $a = 1,\ b_i = 0$, and $a = 0,\ b_i = 1,\ b_{j\neq i} = 0$ respectively, we have

$$\mathbb{E}[\mathbb{L}[X \mid Y] - X] = 0 \implies \mathbb{E}[\mathbb{L}[X \mid Y]] = \mathbb{E}[X]$$
$$\mathbb{E}[(\mathbb{L}[X \mid Y] - X)Y_i] = 0 \implies \mathbb{E}[\mathbb{L}[X \mid Y]Y_i] = \mathbb{E}[XY_i]$$

The first equation says that the best linear estimator is unbiased. The second equation says that the observation error is orthogonal to each observation; i.e. the observation error is uncorrelated the observations.

This is called the "orthogonality principle". It uniquely characterizes $\mathbb{L}[X \mid Y]$. □

To see if this agrees with what we derived before, try plugging in

$$\mathbb{L}[X \mid Y] = \mu_X + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}{}^{-1} Y.$$

We'd get

$$\begin{aligned}\mathbb{E}[\mathbb{L}[X \mid Y]] &= \mu_X = \mathbb{E}[X] \\ \mathbb{E}[\mathbb{L}[X \mid Y]Y_i] &= \mu_X \mathbb{E}[Y_i] + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}{}^{-1}\mathbb{E}\left[(Y - \mu_Y)Y_i\right] \\ &= \mu_X + \Sigma_{\mathbf{XY}} = \mathbb{E}\left[XY^T\right]\end{aligned}$$

4/20/2021

Lecture 23

More Linear Estimation, Online Estimation

Last time, one of the big points was that a collection of RVs with finite second moments is a Hilbert Space (a vector space with an inner product) equipped with the inner product $\langle X, Y\rangle := \mathbb{E}[XY]$. This means that $\|X\| = \sqrt{\mathbb{E}\left[|X|^2\right]}$. This is often denoted as $L^2(\Omega, \mathcal{F}, P)$.

Another way to think about this inner product is $\mathbb{E}[XY] = \sum_{\omega\in\Omega} p(\omega)X(\omega)Y(\omega)$, but it's not too important to dwell on.

One question is why do we focus on squared loss? Because of Hilbert spaces. Hilbert spaces are very structured, and we know a lot about them and can say a lot about them. These niceties are not present with other kinds of losses.

Recall the Hilbert Projection Theorem (Theorem 22.3); this allows us to connect the idea of Hilbert Spaces to linear estimation.

Taking $\mathcal{U} = \{a + \sum_{i=1}^n b_i Y_i : a, b_1, \ldots, b_n \in \mathbb{R}\}$, a subspace of the Hilbert space of RVs with finite second moments, this is the set of all linear estimators of X given the observations $Y_1, Y_2, \ldots, Y_n$. We have (by Definition 22.1)

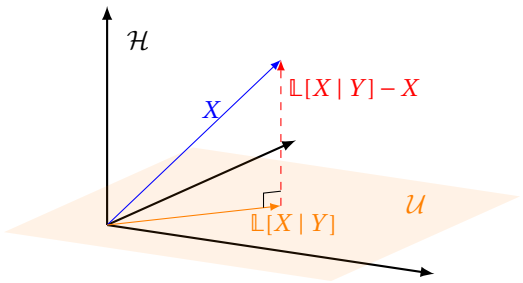
$$\mathbb{L}[X \mid Y] := \operatorname*{argmin}_{U\in\mathcal{U}} \|X - U\|^2 = \operatorname*{argmin}_{\text{linear } \hat{X}} \mathbb{E}\left[\left|X - \hat{X}(Y)\right|^2\right].$$

The projection theorem tells us that this exists and is unique. The orthogonality principle (Theorem 22.5) tells us that the LLSE is unbiased, and that the LLS error is orthogonal to the observations. The latter also implies that the LLS error is uncorrelated with the observations (if the RVs are of zero mean).

It turns out that setting $\mathbb{L}[X \mid Y] = a + \sum b_i Y_i$ and plugging into the orthogonality principle and solving gives us system of linear equations, resulting in

$$\mathbb{L}[X \mid Y] = \mu_X + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1}(Y - \mu_Y).$$

A question we'll tackle today is: what is the error attained by the best linear estimator?



The error of the LLSE is

$$\begin{aligned}
\mathbb{E}\left[|\mathbb{L}[X \mid Y] - X|^2\right] &= \|\mathbb{L}[X \mid Y] - X\|^2 \\
&= \|X\|^2 - \|\mathbb{L}[X \mid Y]\|^2 && \text{(Pythagorean Theorem)} \\
&= \mathbb{E}\left[|X|^2\right] - \mathbb{E}\left[|\mathbb{L}[Y \mid X]|^2\right] \\
&= \mathrm{Var}(X) - \mathbb{E}\left[\left|\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1}(Y - \mu_Y)\right|^2\right] && \text{(assuming zero mean)} \\
&= \mathrm{Var}(X) - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1}\underbrace{\Sigma_{\mathbf{XY}}^T}_{\Sigma_{\mathbf{YX}}}
\end{aligned}$$

In a nutshell, the theory of linear least squared estimation is

$$\begin{aligned}
\mathbb{L}[X \mid Y] &= \mu_Y + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1}(Y - \mu_Y) \\
\text{LLS error} &= \mathrm{Var}(X) - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{YX}}
\end{aligned}$$

23.1 Orthogonality Principle in MMSE

Here's another application of the orthogonality principle.

Let X, Y be RVs with $\mathbb{E}\left[X^2\right] < \infty$. By calculus (see HW),

$$\mathbb{E}[X \mid Y] = \operatorname*{argmin}_{\hat{X}(Y)} \mathbb{E}\left[\left|X - \hat{X}(Y)\right|^2\right].$$

Recall that I told you the "real" definition of conditional expectation was the tower property:

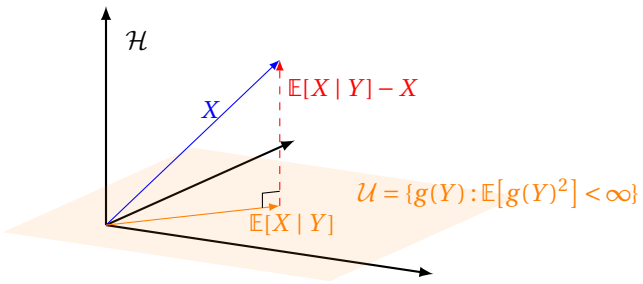
$$\mathbb{E}\left[\mathbb{E}[X \mid Y]g(Y)\right] = \mathbb{E}\left[Xg(Y)\right] \qquad \forall g(Y)$$

assuming $\mathbb{E}\left[g(Y)^2\right] < \infty$.

The orthogonality principle characterization of $\mathbb{E}[X \mid Y]$ is precisely

$$\begin{aligned}
\mathbb{E}\left[(\mathbb{E}[X \mid Y] - X)g(Y)\right] &= 0 \qquad \forall g(Y) \\
\mathbb{E}\left[\mathbb{E}[X \mid Y]g(Y)\right] &= \mathbb{E}\left[Xg(Y)\right] \qquad \forall g(Y)
\end{aligned}$$

That is, the conditional expectation is the projection of X onto the subspace of functions of Y, and the tower property is the same as the characterization by the orthogonality principle.



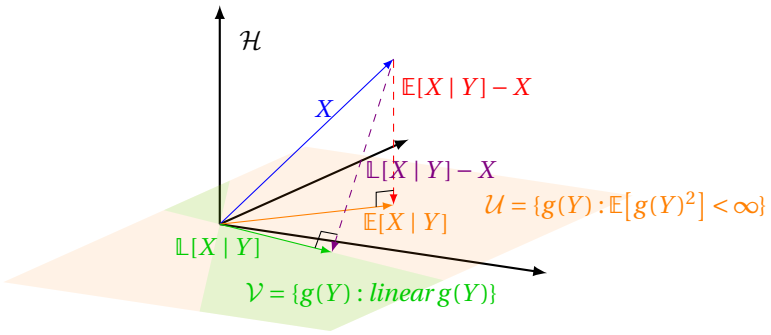
Posing a similar question as before, what is the MMS error? Doing Pythagorean theorem again, assuming zero mean, we have

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

This is just the law of total variance!

Before, when we first introduced conditional expectation and conditional variance, we were very formulaic—but here, we see that these definitions can be thought of as entirely geometric in nature.

A final remark regarding the MMSE vs LLSE: in general, $\mathbb{L}[X \mid Y] \neq \mathbb{E}[X \mid Y]$. The subspace of linear functions ($\mathcal{V}$ below) is a subset of the subspace of all possible functions of Y.



However, there are special cases where $\mathbb{L}[X \mid Y] = \mathbb{E}[X \mid Y]$. Most notably, when X, Y are jointly Gaussian.

23.2 "Online" estimation

Suppose data is observed sequentially. How do we efficiently update our estimate on arrival of new observations?

For motivation, let us start with a simple setting. Let us assume $\mathbb{E}[X] = 0$, and suppose observations $Y_1, Y_2, \ldots$ are orthogonal: $\langle Y_i, Y_j \rangle = 0$ for $i \neq j$. Let us define $Y^n = (Y_1, \ldots, Y_n)$ for convenience.

Our claim is that $\mathbb{L}\left[X \mid Y^{n+1}\right] = \mathbb{L}[X \mid Y^n] + \mathbb{L}[X \mid Y_{n+1}]$. In this case,

$$\mathbb{L}\left[X \mid Y^{n+1}\right] = \mathbb{L}\left[X \mid Y^n\right] + \frac{\text{Cov}(X, Y_{n+1})}{\text{Var}(Y_{n+1})}(Y_{n+1} - \mu_{n+1}).$$

What we can see here is that we can update our best estimate by just adding a term.

Proof. All we need to do is check the orthogonality principle.

$$\mathbb{E}\left[\left(\mathbb{L}\left[X \mid Y^{n+1}\right] - X\right)Y_k\right] = \mathbb{E}\left[\left(\mathbb{L}\left[X \mid Y^n\right] + \mathbb{L}[X \mid Y_{n+1}] - X\right)Y_k\right].$$

If $k = n+1$, we have

$$\mathbb{E}\left[\left(\mathbb{L}\left[X \mid Y^n\right] + \mathbb{L}[X \mid Y_{n+1}] - X\right)Y_{n+1}\right] = \mathbb{E}[(\mathbb{L}[X \mid Y_{n+1}] - X)Y_{n+1}] + \mathbb{E}\left[\mathbb{L}\left[X \mid Y^n\right]Y_{n+1}\right].$$

The first term is 0 by the orthogonality principle (because $\mathbb{L}[X \mid Y_{n+1}] - X$ is the error of $\mathbb{L}[X \mid Y_{n+1}]$), and the second term is also 0 since $\mathbb{L}[X \mid Y^n]$ is a linear function of $Y_1, \dots, Y_n$ and we've assumed that all Y_i's are orthogonal.

If $k \le n$, we'd just expand the expectation the other way:

$$\mathbb{E}\left[\left(\mathbb{L}\left[X \mid Y^n\right] + \mathbb{L}[X \mid Y_{n+1}] - X\right) Y_{n+1}\right] = \mathbb{E}\left[\left(\mathbb{L}\left[X \mid Y^n\right] - X\right) Y_k\right] + \mathbb{E}[\mathbb{L}[X \mid Y_{n+1}] Y_k].$$

The first term is 0 by the orthogonality principle for $\mathbb{L}[X \mid Y^n]$, and the second term is also 0 by our assumptions, as $\mathbb{L}[X \mid Y_{n+1}]$ is a linear function of Y_{n+1}. □

4/22/2021

Lecture 24

Gram-Schmidt, Jointly Gaussian RVs

Continuing on from last time, we have the following update procedure:

Theorem 24.1: Updating the LLSE

If Y_i's are uncorrelated with $Y_0 = 0$, we have

$$\mathbb{L}[X \mid Y_0] = \mathbb{E}[X],$$

and for $n = 0, 1, 2, \dots$, we have

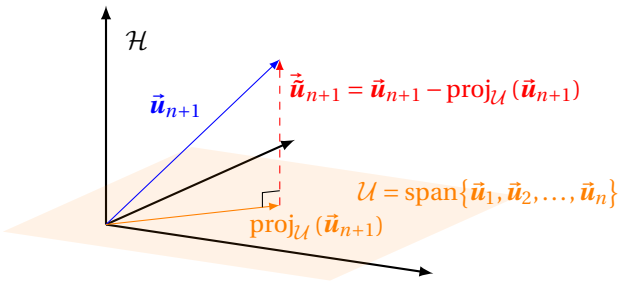
$$\mathbb{L}\left[X \mid Y^{n+1}\right] = \mathbb{L}\left[X \mid Y^n\right] + \frac{\operatorname{Cov}(X, Y_{n+1})}{\operatorname{Var}(Y_{n+1})}(Y_{n+1} - \mathbb{E}[Y_{n+1}]).$$

Here, Y^n denotes the sequence $(Y_1, \dots, Y_n)$.

Hence, if our observations are uncorrelated, we have a ridiculously nice way of sequentially updating our estimate of X given new observations.

In general though, observations are not uncorrelated. But, if our observations are correlated, we can transform them to be uncorrelated using Gram-Schmidt.

In linear algebra, we can take non-orthogonal vectors $\vec{u}_1, \vec{u}_2, \dots$, and make them orthogonal by running Gram-Schmidt. Through a picture, we have



The resulting sequence $\tilde{\vec{u}}_1, \tilde{\vec{u}}_2, \dots$ are orthogonal, and $\text{span}\left\{\tilde{\vec{u}}_1, \tilde{\vec{u}}_2, \dots, \tilde{\vec{u}}_n\right\} = \text{span}\{\vec{u}_1, \vec{u}_2, \dots, \vec{u}_n\}$ for all $n \ge 1$.

Definition 24.2: Gram-Schmidt for Random Variables

Given a sequence of RV's $Y_1, Y_2, \dots$, define $\tilde{Y}_{n+1} = Y_{n+1} - \mathbb{L}[Y_{n+1} \mid Y^n]$.

The result is that $\tilde{Y}_1, \tilde{Y}_2, \dots$ are uncorrelated by the Gram-Schmidt construction (which is really just the orthogonality principle). Furthermore, $\text{span}\{1, \tilde{Y}_1, \dots, \tilde{Y}_n\} = \text{span}\{1, Y_1, \dots, Y_n\}$ for $n \ge 1$.

What this means is that

$$\mathbb{L}\big[X \mid Y^n\big] = \mathbb{L}\big[X \mid \tilde{Y}^n\big] \qquad \forall n \geq 1.$$

The key thing gained here is that the $\tilde{Y}^n$'s are uncorrelated, so we can sequentially compute $\mathbb{L}\big[X \mid \tilde{Y}^n\big]$, and therefore $\mathbb{L}[X \mid Y^n]$.

The sequence $\tilde{Y}_1, \tilde{Y}_2, \ldots$ is called the linear innovation sequence (sometimes also called the orthogonal innovations) corresponding to $Y_1, Y_2, \ldots$.

$\tilde{Y}_n$ is the component of Y_n that can't be linearly estimated from $Y_1, \ldots, Y_n$ (equivalently from $\tilde{Y}_1, \ldots, \tilde{Y}_n$.

So, conceptually, by transforming $Y_1, Y_2, \ldots \to \tilde{Y}_1, \tilde{Y}_2, \ldots$, we can always do sequential updates of our linear estimator using the ridiculously simple update we saw earlier.

This is the big idea behind the Kalman Filter. We'll come to this soon.

24.1 Jointly Gaussian Random Variables

Definition 24.3: Jointly Gaussian Random Variables

A gaussian random vector $X = (X_1, \ldots, X_n)^T$, i.e. jointly gaussian random variables, with density on $\mathbb{R}^n$ is defined via its pdf:

$$f_X(x) = \frac{1}{(2\pi)^{\frac{n}{2}} \det(\Sigma_{\mathbf{X}})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_X)^T \Sigma_{\mathbf{X}}{}^{-1}(x-\mu_X)\right).$$

Here, $\Sigma_{\mathbf{X}} = \mathrm{Cov}(X) = \mathbb{E}\big[(X-\mu_X)(X-\mu_X)^T\big]$, and $\mu_X = \mathbb{E}[X] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}$.

We write $X \sim \mathcal{N}(\mu_X, \Sigma_{\mathbf{X}})$ for short.

One observation here is that Gaussian vectors have distributions parameterized entirely by mean and covariance (similar to gaussians in one dimension).

There are many equivalent definitions of gaussian vectors.

1. We can define them via the pdf as above.
2. Gaussian random vectors are affine transformations of iid gaussian random variables.

   That is, if X has a nonsingular $\Sigma_{\mathbf{X}}$, then we can write

   $$X = \mu_X + \mathbf{A}W,$$

   where $\mathbf{A} \in \mathbb{R}^{n \times n}$ with full rank, and $W = (W_1, \ldots, W_n)^T$, with $W_i \sim_{iid} \mathcal{N}(0,1)$.

   Why is this equivalent to (1)? We can use derived distributions. Namely, we have (here, $|\cdot|$ on a vector represents the euclidean distance in $\mathbb{R}^n$):

   $$\begin{aligned} f_X(x) &= \frac{1}{|\det(\mathbf{A})|} f_W\big(\mathbf{A}^{-1}(x-\mu_X)\big) \\ &= \frac{1}{|\det(\mathbf{A})|} \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\big|\mathbf{A}^{-1}(x-\mu_X)\big|^2\right) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} \det\big(\mathbf{A}\mathbf{A}^T\big)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_X)^T(\mathbf{A}\mathbf{A}^T)^{-1}(x-\mu_X)\right) \end{aligned}$$

   But, we have

   $$\mathrm{Cov}(X) = \mathbb{E}\big[(X-\mu_X)(X-\mu_X)^T\big] = \mathbb{E}\big[\mathbf{A}WW^T\mathbf{A}^T\big] = \mathbf{A}\mathbf{I}\mathbf{A}^T = \mathbf{A}\mathbf{A}^T.$$

   Hence, this agrees with the density definition.

3. X is a gaussian random vector iff all of its one-dimensional projections are gaussian random variables; i.e.
$$a^T X \sim \mathcal{N}(a^T \mu_X, a^T \Sigma_{\mathbf{X}} a) \qquad \forall a \in \mathbb{R}^n.$$

For random vectors X, Y, we can partition the covariance matrix as
$$\mathrm{Cov}\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right) = \begin{bmatrix} \mathbb{E}\big[(X-\mu_X)(X-\mu_X)^T\big] & \mathbb{E}\big[(X-\mu_X)(Y-\mu_Y)\big] \\ \mathbb{E}\big[(Y-\mu_Y)(X-\mu_X)^T\big] & \mathbb{E}\big[(Y-\mu_Y)(Y-\mu_Y)\big] \end{bmatrix} = \begin{bmatrix} \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{Y}} \end{bmatrix}.$$

An amazing fact is that if X, Y are jointly gaussian vectors (i.e. $\begin{bmatrix} X \\ Y \end{bmatrix}$ is a gaussian vector), then we can always write
$$X = \mu_X + \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} (Y - \mu_Y) + V,$$
where $V \sim \mathcal{N}(0, \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{YX}})$ independent of Y. Notice here that the first part is the $\mathbb{L}[X \mid Y]$.

Proof. Let $\tilde{X} = \mu_X + \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} (Y - \mu_Y) + V$. Our claim is that $\tilde{X}$, Y are jointly gaussian.

Since Y and V are independent gaussians, we can write them as $Y = \mu_Y + \mathbf{A} W_1$ for some $\mathbf{A}$ with $W_1 \sim \mathcal{N}(0, \mathbf{I})$, and $V = \mathbf{B} W_2$ for some $\mathbf{B}$ with $W_2 \sim \mathcal{N}(0, \mathbf{I})$ independent of W_1.

By definition, we can write
$$\begin{bmatrix} \tilde{X} \\ Y \end{bmatrix} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} + \begin{bmatrix} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \mathbf{A} & \mathbf{B} \\ \mathbf{A} & 0 \end{bmatrix} \begin{bmatrix} W_1 \\ W_2 \end{bmatrix}.$$
This is just an affine transformation of iid standard gaussians, so $\tilde{X}$ and Y is also jointly gaussian.

Since $(\tilde{X}, Y)$ is jointly gaussian, its distribution is parameterized entirely by mean and covariance.

We have $\mathbb{E}\big[\tilde{X}\big] = \mu_X$, and $\mathbb{E}[Y] = \mu_Y$. We also have
$$\begin{aligned}
\Sigma_{\tilde{\mathbf{X}}\mathbf{Y}} &= \mathbb{E}\big[(\tilde{X} - \mu_X)(Y - \mu_Y)^T\big] \\
&= \mathbb{E}\big[\big(\Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} (Y - \mu_Y) + V\big)(Y - \mu_Y)^T\big] \\
&= \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \mathbb{E}\big[(Y - \mu_Y)(Y - \mu_Y)^T\big] \\
&= \Sigma_{\mathbf{XY}} \\
\Sigma_{\tilde{\mathbf{X}}} &= \mathbb{E}\big[(\tilde{X} - \mu_X)(\tilde{X} - \mu_X)^T\big] \\
&= \cdots \\
&= \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \mathbb{E}\big[(Y - \mu_Y)(Y - \mu_Y)^T\big] \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{YX}} \\
&= \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{YX}} + \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{YX}} \\
&= \Sigma_{\mathbf{X}}
\end{aligned}$$
So, $(\tilde{X}, Y)$ is jointly gaussian, with mean $\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$ and covariance $\begin{bmatrix} \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{Y}} \end{bmatrix}$.

Since this is the same parameterization as (X, Y), so it must be equal in distribution, and it must be the case that (X, Y) are also jointly gaussian. □

Theorem 24.4: MMSE for gaussians

A corollary of this is that for X, Y jointly gaussian,
$$\mathbb{E}[X \mid Y] = \mathbb{E}\big[\mu_X + \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} (Y - \mu_Y) + V \mid Y\big] = \mu_X + \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} (Y - \mu_Y) = \mathbb{L}[X \mid Y].$$

This tells us that linear least squares estimation coincides with the (optimal) minimum mean square error estimation for gaussians.

In practice, things are often approximately gaussian (by CLT). This means that we can expect linear estimation in these instances to perform near-optimally.

4/27/2021

Lecture 25

Kalman Filter

From last time, we saw that there are many equivalent characterizations of jointly gaussian random variables:

1. Specify the density
2. Affine transformations of iid $\mathcal{N}(0,1)$ RVs
3. All 1-dim projections are gaussian RVs
4. maximum entropy distribution subject to second moment constraints
5. limit in CLT
6. etc.

The most important property (arguable) is the following. If X, Y are jointly gaussian random vectors, then we can write

$$X = \mu_X + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}{}^{-1}\big(Y - \mu_Y\big) + V,$$

where $V \sim \mathcal{N}(0, \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}{}^{-1}\Sigma_{\mathbf{YX}})$. That is, in a gaussian vector, each coordinate is a "noisy" version of the others (i.e. X is a linear transformation of Y, plus some noise).

A consequence of this is that

$$\mathbb{L}[X \mid Y] = \mathbb{E}[X \mid Y] = \mu_X + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}{}^{-1}\big(Y - \mu_Y\big).$$

That is, linear estimation is optimal for gaussians.

One caution is that gaussian marginals does not imply jointly gaussian.

Example 25.1

Take $Y \sim \mathcal{N}(0,1)$, and let $B = \begin{cases} +1 & \text{w.p. } \frac{1}{2} \\ -1 & \text{w.p. } \frac{1}{2} \end{cases}$ be independent of Y.

Let $X = BY \sim \mathcal{N}(0,1)$. This means that X and Y are both $\mathcal{N}(0,1)$.

But, (X,Y) are not jointly gaussian; the mass of (X,Y) are concentrated on an X shape about the origin.

25.1 Kalman Filter

We already did all of the heavy lifting (i.e. the theory is all done). Today is mostly plug and chug of what we already know.

The basic setting is a state space model. Suppose we have $X_0, V_0, V_1, \ldots, W_0, W_1, \ldots$ be uncorrelated random vectors, say with zero mean (WLOG).

A state space model contains an evolution of the form

$$X_{n+1} = \mathbf{A}X_n + V_n,$$

where $n \geq 0$ and $\mathbf{A}$ is a matrix.

We also have observations of the form

$$Y_n = \mathbf{C}X_n + W_n,$$

where $n \geq 1$ and $\mathbf{C}$ is a matrix.

This is a flexible model for a variety of processes. Note that if $X_0, V_0, V_1, \ldots, W_0, W_1, \ldots$ are gaussians, then everything is jointly gaussian.

Example 25.2

Let $p(n)$ be the position at time n. Let

$$X_n = \begin{bmatrix} p(n) \\ p(n-1) \\ p(n-2) \end{bmatrix}.$$

We then have

$$X_{n+1} = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} X_n + \begin{bmatrix} Z_n \\ 0 \\ 0 \end{bmatrix},$$

where $Z_n \sim \mathcal{N}(0, \sigma^2)$. We also have

$$Y_n = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} X_n + W_n,$$

where $W_n \sim \mathcal{N}(0, \zeta^2)$.

A Kalman Filter is an efficient algorithm for estimating X process sequentially from the observations.

Many variations are possible (by choosing $\mathbf{A}$ and $\mathbf{C}$):

1. Prediction: estimate X_{n+k} from $Y_1, \ldots, Y_n$.
2. Filtering: estimate X_n from $Y_1, \ldots, Y_n$.
3. Smoothing: estimate X_{n-k} from $Y_1, \ldots, Y_n$.

Theorem 25.3: Kalman Filter

Let $(X_n)_{n \geq 0}$ evolve according to the state space model above. Let $\hat{X}_{n|m}$ denote $\mathbb{L}[X_n \mid Y^m]$, and let $\Sigma_{n|m}$ denote $\text{Cov}(X_n - \hat{X}_{n|m}$, which is the covariance of the estimation error of X_n given Y^m. Further, let $\Sigma_{\mathbf{V}} = \text{Cov}(V - i)$ and $\Sigma_{\mathbf{W}} = \text{Cov}(W_i)$ for $i \geq 0$ (these can change with time, but we will assume that they are all equal for all i).

Initialize $\hat{X}_{0|0} = 0$, and $\Sigma_{\mathbf{0|0}} = \text{Cov}(X_0)$.

For $n \geq 1$, do

$$\begin{aligned} \hat{X}_{n|n} &= \mathbf{A}\hat{X}_{n-1|n-1} + \mathbf{K}_n\left(Y_n - \mathbf{CA}\hat{X}_{n-1|n-1}\right) \\ \mathbf{K}_n &= \Sigma_{n|n-1}\mathbf{C}^T\left(\mathbf{C}\Sigma_{n|n-1}\mathbf{C}^T + \Sigma_{\mathbf{W}}\right)^{-1} \\ \Sigma_{n|n-1} &= \mathbf{A}\Sigma_{n-1|n-1} + \Sigma_{\mathbf{V}} \\ \Sigma_{n|n} &= (\mathbf{I} - \mathbf{K}_n\mathbf{C})\Sigma_{n|n-1} \end{aligned}$$

Proof. We'll verify the correctness for the scalar case of the Kalman Filter. Let $X_n = aX_{n-1} + V_n$, $Y_n = X_n + W_n$ for $n \geq 1$. Let X_0 have zero mean.

We initialize $\hat{X}_{0|0} = 0$, and $\sigma^2_{0|0} = \text{Var}(X_0)$.

The updates are:

$$\begin{aligned} X_{n|n} &= a\hat{X}_{n-1|n-1} + K_n\left(Y_n - a\hat{X}_{n-1|n-1}\right) \\ K_n &= \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_W} \\ \sigma^2_{n|n-1} &= a^2\sigma^2_{n-1|n-1} + \sigma^2_W \\ \sigma^2_{n|n} &= (1 - K_n)\sigma^2_{n|n-1} \end{aligned}$$

The proof of correctness for the scalar Kalman Filter is with induction:

The initialization is trivially correct. So, it suffices to verify the update equations.

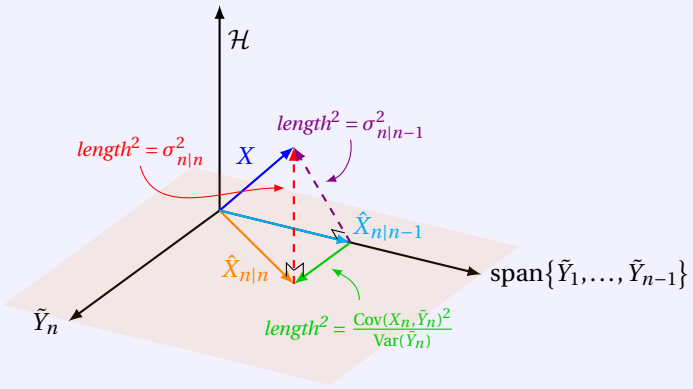
Let $\tilde{Y}_1, \tilde{Y}_2, \ldots$ be the linear innovation sequence corresponding to $Y_1, Y_2, \ldots$.

We then have

$$\begin{aligned}
\hat{X}_{n|n} &= \mathbb{L}\left[X_n \mid Y^n\right] \\
&= \mathbb{L}\left[X_n \mid \tilde{Y}^n\right] \\
&= \mathbb{L}\left[X_n \mid \tilde{Y}^{n-1}\right] + \mathbb{L}\left[X_n \mid \tilde{Y}^n\right] \\
&= \hat{X}_{n|n-1} + \underbrace{\frac{\mathrm{Cov}(X_n, \tilde{Y}_n)}{\mathrm{Var}(\tilde{Y}_n)}}_{:=K_n} \tilde{Y}_n
\end{aligned}$$

By properties of error for LLSE, we have

$$\begin{aligned}
\sigma^2_{n|n} &= \sigma^2_{n|n-1} - \frac{\mathrm{Cov}(X_n, \tilde{Y}_n)^2}{\mathrm{Var}(\tilde{Y}_n)} \\
&= \sigma^2_{n|n-1} - K_n \,\mathrm{Cov}(X_n, \tilde{Y}_n)
\end{aligned}$$



Now, let's evaluate

$$\begin{aligned}
\mathrm{Cov}(X_n, \tilde{Y}_N) &= \mathrm{Cov}(X_n, Y_n - \mathbb{L}\left[Y_n \mid \tilde{Y}^{n-1}\right]) \\
&= \mathrm{Cov}(X_n, Y_n - \mathbb{L}\left[X_n + W_n \mid \tilde{Y}^{n-1}\right]) \\
&= \mathrm{Cov}(X_n, Y_n - \mathbb{L}\left[X_n \mid \tilde{Y}^{n-1}\right]) && W_n \text{ uncorrelated} \\
&= \mathrm{Cov}(X_n, X_n + W_n - \mathbb{L}\left[X_n \mid \tilde{Y}^{n-1}\right]) && W_n \text{ uncorrelated} \\
&= \mathrm{Cov}(X_n, X_n - \mathbb{L}\left[X_n \mid \tilde{Y}^{n-1}\right]) \\
&= \mathrm{Cov}(X_n, X_n - \mathbb{L}\left[X_n \mid \tilde{Y}^{n-1}\right]) + \underbrace{\mathrm{Cov}(\mathbb{L}\left[X_n \mid \tilde{Y}^{n-1}\right], X_n - \mathbb{L}\left[X_n \mid \tilde{Y}^{n-1}\right])}_{0 \text{ by orthogonality principle}} && \text{adding } 0 \\
&= \mathrm{Cov}(X_n - \mathbb{L}\left[X_n \mid \tilde{Y}^{n-1}\right], X_n - \mathbb{L}\left[X_n \mid \tilde{Y}^{n-1}\right]) \\
&= \mathrm{Var}(X_n - \mathbb{L}\left[X_n \mid \tilde{Y}^{n-1}\right]) \\
&= \sigma^2_{n|n-1}
\end{aligned}$$

Note that

$$\begin{aligned}\hat{X}_{n|n} &= \hat{X}_{n|n-1} + K_n\big(Y^n - \mathbb{L}\big[Y_n \mid \tilde{Y}^{n-1}\big]\big)\\ &= \mathbb{L}\big[aX_n + \cancel{V_{n-1}} \mid \tilde{Y}^{n-1}\big] + K_n(Y_n - \underbrace{\mathbb{L}\big[X_n + \cancel{W_n} \mid \tilde{Y}^{n-1}\big]}_{\hat{X}_{n|n-1}})\\ &= a\mathbb{L}\big[X_{n-1} \mid \tilde{Y}^{n-1}\big] + K_n\big(Y_n - a\mathbb{L}\big[X_{n-1} \mid \tilde{Y}^{n-1}\big]\big)\\ &= a\hat{X}_{n-1|n-1} + K_n\big(Y_n - a\hat{X}_{n-1|n-1}\big)\end{aligned}$$

We also have

$$\begin{aligned}\sigma^2_{n|n} &= \sigma^2 - K_n \operatorname{Cov}(X_n, \tilde{Y}^n)\\ &= (1-K_n)\sigma^2_{n|n-1}\end{aligned}$$

This verifies the update of $\sigma^2_{n|n}$.

We also have

$$\begin{aligned}\sigma^2_{n|n-1} &= \mathbb{E}\Big[X_n - \mathbb{L}\big[X_n \mid \tilde{Y}^{n-1}\big]^2\Big]\\ &= \mathbb{E}\Big[\big(aX_{n-1} + V_n - \mathbb{L}\big[aX_{n-1} + V_n \mid \tilde{Y}^{n-1}\big]\big)^2\Big]\\ &= a\sigma^2_{n-1|n-1} + \sigma^2_V\end{aligned}$$

This verifies the update of $\sigma^2_{n|n-1}$.

We already have $K_n = \frac{\sigma^2_{n|n-1}}{\operatorname{Var}(\tilde{Y}_n)}$.

We then have to evaluate the variance:

$$\begin{aligned}\operatorname{Var}(\tilde{Y}_n) &= \mathbb{E}\Big[\big(Y - n - \mathbb{L}\big[Y_n \mid \tilde{Y}^{n-1}\big]\big)^2\Big]\\ &= \mathbb{E}\Big[\big(X_n + W_n - \mathbb{L}\big[X_n + \cancel{W_n} \mid \tilde{Y}^{n-1}\big]\big)^2\Big]\\ &= \sigma^2_{n|n-1} + \sigma^2_W\end{aligned}$$

This verifies the update of K_n. □